FAST SAMPLING RATE CONSIDERATIONS IN DIGITAL CONTROL DESIGN : CYCLOSTATIONARY AND FINITE WORDLENGTH ISSUES

Kusmayanto Kadiman

February 1988

A thesis submitted for the degree of

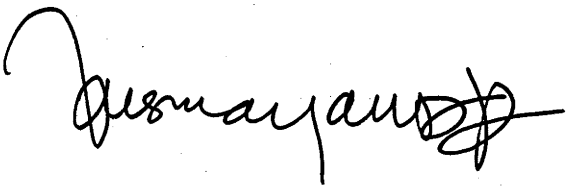
DOCTOR OF PHILOSOPHY

of the Australian National University

The doctoral studies were conducted under the supervision of Dr. D. Williamson and the advice of Professor J.B. Moore and Dr. R.R. Bitmead.

"I hereby declare that the result presented in this thesis, except as otherwise explicitly stated is based on personal and original efforts and has not been submitted for any other degree to any university or other educational institutions"

Canberra, February 1988



Kusmayanto Kadiman
Department of Systems Engineering
Research School of Physical Sciences
The Australian National University
GPO Box 4 ACT 2601
Canberra, Australia

Bismillahirrahmanirrahim

ACKNOWLEDGEMENT

Technically, I would like to thank Dr. Darrell Williamson for his guidance and supervision during my Doctoral studies in the Department of Systems Engineering, the Australian National University and partially at the Systems and Control Department, the University of New South Wales.

I believe that it is impossible for me to satisfactorily thank Srie, Nuki, Tantri and Didong for their support in particular for allowing me to use their valuable time.

Special tribute also goes to the staff and students of the Department of Systems Engineering, Research School of Physical Sciences for providing a warm and highly stimulating (research) environment.

I thank also The Australian International Development Assistance Bureau for providing the scholarship. I should also mention the Systems and Control Department, the University of New South Wales, Australia for providing the opportunity to conduct my first year of studies in Australia.

ABSTRACT

Cyclostationarity and *finite wordlength* issues which are particularly relevant for digital control systems operating under a *fast* sampling rate are considered in this thesis.

The properties of some cyclostationary processes which may result as a consequence of implementing a digital controller are investigated. A state space translational representation of a cyclostationary process is developed and is used to characterize two classes of first and second order processes. A first order approximation of a harmonic series representation is also examined.

A minimax quadratic cost function which takes into account the periodic nature of the statistics is defined and optimized. The resulting linear state feedback law improves the intersample behavior of the digitally controlled systems. In particular, the minimax variance regulator which is proposed as a new design method for regulating the output variance offers a significant improvement over the classical minimum variance regulator.

A conventional optimal state prediction algorithm is developed under the assumption that the system is driven by (wide sense) stationary disturbances. The consequences of cyclostationary disturbances on the optimal state estimation is also explored in this thesis. For a single rate case (ie. equal control and output sampling rates), it is shown that synchronous measurement and control is not necessarily optimal. A multirate state predictor is designed to produce the optimal state vector prediction based on some N measurements where N is the ratio between the output and control sampling rates.

In practice, when sampling rates are high or on-board facilities are limited, low precision algorithms must be implemented. In this thesis, the finite wordlength constraint is directly incorporated into the design of a linear quadratic Gaussian regulator. The optimal finite state wordlength regulator problem involves the problem

of finding the optimal Kalman filter and controller gains, and of selecting the optimal structure. An iterative procedure is proposed for solving these problems.

The finite coefficient wordlength consideration in the implementation of a linear quadratic regulator is also discussed. The integer residue correction schemes is incorporated in the design.

In terms of structural complexity, the optimum structure is inappropriate for implementation since in general all coefficients are neither zero nor one. A delay replaced direct form (DRDF) structure is proposed instead. The DRDF structure is a good candidate since it has low complexity structure and low sensitivity to coefficients change due to a finite coefficient wordlength implementation. The DRDF structure can be derived directly from a given default structure by using the procedures which are developed in this thesis. The procedures are developed for both single-input single-ouput and multi-input multi-output systems.

TABLE OF CONTENTS

Acknowledgement

Abstract

List of figures and tables

1. INTRODUCTION
   1.1 Motivation 1
   1.2 Cyclostationarity in the digital regulation of continuous-time systems : Concepts and consequences 4
   1.3 Finite wordlength effects in digital control applications : An overview 8
   1.4 Research areas and organization of the thesis 23
   1.5 Original ideas proposed in the thesis 26

2. CYCLOSTATIONARY PROCESSES
   2.1 Introduction 28
   2.2 Discretization of continuous-time systems 30
   2.3 Representation of cyclostationary processes 40
   2.4 Relevant cyclostationary processes in stochastic control 49
   2.5 Conclusions 61

3. OPTIMAL STATE PREDICTION
   3.1 Introduction 63
   3.2 Non-synchronous state prediction 67
   3.3 Multirate state prediction 75
   3.4 Conclusions 82

4. DISCRETE LINEAR QUADRATIC REGULATION OF CONTINUOUS-TIME SYSTEMS
   4.1 Introduction 84
   4.2 Linear quadratic Gaussian regulator design 85
   4.3 Minimax quadratic regulation 93
   4.4 Minimax output variance regulation 103
   4.5 Conclusions 118

5. FINITE WORDLENGTH LINEAR QUADRATIC REGULATOR DESIGN
5.1 Introduction 120
5.2 Finite wordlength linear quadratic Gaussian regulator 127
5.3 Optimum gains : Default structure 135
5.4 Optimum structure : Fixed gains 150
5.5 Optimum FSWL-LQG regulator 159
5.6 Coefficient wordlength consideration in LQG regulator 161
5.7 Conclusions 173

6. A LOW COMPLEXITY - LOW SENSITIVITY COMPENSATOR IMPLEMENTATION
6.1 Introduction 175
6.2 Single-input single-output delay replaced direct form structures 179
6.3 Multi-input multi-output delay replaced direct form structures 192
6.4 Conclusions 204

7. SUMMARY, CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH 206

APPENDICES
Appendix A : Longitudinal control of a modern transport aeroplane 211
Appendix B : An Euler-Bernoulli beam model 217

PUBLICATIONS ARISING FROM THE RESEARCH 227

REFERENCES 228

LIST OF FIGURES AND TABLES

FIGURES

Fig.2.1 Discrete equivalent of a continuous-time system 32

Fig.2.2 The stability regions of continuous-time and the corresponding discrete-time models. 34

Fig.2.3 Covariance $\Omega(\delta)$ for 1st-order systems 43

Fig.2.4 Covariance $\Omega(\delta)$ for 2nd-order systems 45

Fig.2.5 Coefficients of covariance approximation $\hat{\Omega}(\delta)$ 48

Fig.2.6 Intersample variance $X(n\Delta)$ of a 1st-order system 52

Fig.2.7 Intersample output variance of a 2nd-order system 53

Fig.2.8 Intersample output variance of 2nd-order systems 54

Fig.2.9 A simple example of a cyclostationary disturbance 59

Fig.3.1 Timing diagram of non-synchronous control and output sampling 68

Fig.3.2a Noise covariances V_d and $\tilde{\Omega}_{1-\delta}$ 74

Fig.3.2b Error covariances Q_δ and P_δ 75

Fig.3.3 Timing diagram of measurement and control instants 76

Fig.4.1 The computational delay requirement 89

Fig.4.2 The intersample output variance of a 1st-order system corresponding to
1. MVR (=MMVR)
2. KS 109

Fig.4.3 The intersample output variance of a 2nd-order system corresponding
1. MVR
2. MMVR for $b/a=10^3$ 110

Fig.4.4 The intersample output variance of a 2nd-order system corresponding to
1. MMVR for $b/a=10^3$
2. Cheap control for R=0.015
3. KS 111

Fig.4.5 The maximum output variance versus control weighting R for cheap control in
1. Example 3.2
2. Example 3.3
3. Example 3.4 112

Fig.4.6 The maximum output variance versus control weighting R for MMVR in
1. Example 3.2
2. Example 3.3
3. Example 3.4 113
Fig.4.7 The maximum output variance versus ratio b/a for MMVR in
1. Example 3.2
2. Example 3.3
3. Example 3.4 113
Fig.4.8 The intersample output variance of 2nd–order systems corresponding to the MVR gains for
1. L = [0.1 1]
2. L = [1 1]
3. L = [1 0.1] 114
Fig.4.9 The intersample output variance for L=[1 0.1] corresponding to
1. MVR
2. MMVR for b/a=10^3
3. Cheap control for R=0.015 115
Fig.4.10 The intersample output variance of a 3rd–order system corresponding to
1. Cheap control for R=0.01
2. MMVR for b/a=10^3
3. KS 117
Fig.5.1 Digital control system configuration 129
Fig.6.1 Delay replacement transformation 177
Fig.6.2 2nd order realization of
a. a direct form structure and
b. the corresponding DRDF structure 178
Fig.6.3 A 2–input digital compensator 193
Fig.6.4 4th–order 2–input 2–output compensators realized using
a. direct form structure and
b. the corresponding DRDF structure 199

TABLES

Table 2.1 The effects of sampling on the intersample variance of a 2nd–order system 53
Table 3.1 The influence of the shift factor α on the state prediction error 82
Table 5.1 The effects of the integer residue correction (IRC) on the compensator performance 137

Table 5.2 The compensator performance resulting from
a. Ideal (q=0) design
b. FSWL design 1 149
Table 5.3 The compensator performance resulting from
a. Ideal (q=0) design : default structure
b. FSWL design 2 : Fixed gains (algorithm 5.2) 159
Table 5.4 The compensator performance resulting from
a. Ideal (q=0) design
b. FSWL-LQG design (ie. Algorithm 5.3) 161
Table 5.5 The degradation of the compensator performance due to finite coefficient wordlength for three different compensators described in (5.69a-c) 170
Table 5.6 The FWL performance resulting from
a. Ideal (q=0) design (denoted by $\{K_\infty, G_\infty, T=I\}$)
b. FSWL-LQG design (denoted by {denoted by $\{K_q, G_q, T\}_{opt}$) 172
Table 6.1 The round-off noise performance of compensators implemented using
a. Default structure
b. Controllable scaled DRDF structure 191
Table 6.2 The FWL performance of compensator implemented using
a. Default structure
b. Controllable scaled DRDF structure
c. 'Optimum' FSWL structure
d. Controllable direct form structure 192
Table 6.3 The FWL performance of MIMO compensators implemented using
a. Default structure
b. Controllable scaled DRDF structure
c. Controllable direct form (DF) structure 203

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Since the early sixties, various methods have been developed for designing digital compensators. The early development concerned the translation of analog design; digital PID–controllers [Goff (1966)] and digital re–design (or discretization of continuous–time controllers) [Kuo (1963)] are classical examples. Modern approaches which are inspired by the development of the state–space theory include optimal regulators, pole–placement design, observer theory and optimal filtering [Kwakernaak and Sivan (1972), Anderson and Moore (1979), Franklin and Powell (1980), Åström and Wittenmark (1984)]. Adaptive version of these design have also been investigated in recent years [Goodwin and Sin (1984), Åström and Wittenmark (1984)]. Most digital design techniques have two things in common. First, the attention is often focussed entirely on the sample data response ignoring the intersample behavior. Secondly, the compensators are typically designed assuming the availability of an *infinite* (ie. sufficiently high) *precision* computer system ignoring the fact that 'on–board' facilities are usually limited.

From an analytical point of view, the simplest way of controlling a continuous–time system via a digital computer is to generate the control law as a pulse–amplitude–modulated (PAM) signal with constant period; that is, to produce the continuous control signal u(t) from the sequence $\{u(kT_c)\}$ which is updated at the rate $1/T_c$. At the digital–to–analog conversion times when samples $u(kT_c)$ are converted into the analog signal u(t) (and at the analog–to–digital conversion times when the analog signal y(t) is sampled to produce the sequence $\{y(kT_c)\}$), the closed loop system has a discrete equivalent description. It is well known that the

intersample (ie. between two consecutive sampling instants) response of a continuous-time digitally controlled system can vary significantly from that observed at the sampling instants, particularly when controlling lightly damped open-loop plants using fast sampling rate digital controllers. In deterministic systems, this phenomena is known as the intersample ripple [Tou (1959)]. In stochastic systems, when the systems are driven by wide sense stationary (WSS) [Papoulis (1965)] process disturbance and measurement noise the corresponding discrete process disturbance and measurement noise at the sampling instants $t_k=kT_c$ are also WSS. However, as is shown in [de Souza and Goodwin (1984)], the PAM signal u(t) is not stationary but exhibits *periodic statistical properties*, and consequently so will the system states and outputs. A process which exhibits periodic statistical properties is called *a cyclostationary* process [Bennet (1958)] or equivalently, periodically-stationary, periodic-stationary, periodically-correlated, periodic-nonstationary or periodically-nonstationary [Papoulis (1965), Ogura (1971), Stratonovich (1963), Deutsch (1954), Bendat and Piersol (1966)]. One of the major results of this thesis is the demonstration of the improvement in digital control performance that can be obtained by recognizing the periodic statistical properties of the continuous-time response and by taking the intersample behavior into account.

The increasing availability of microprocessors has stimulated the use of digital controllers [Gupta and Toong (1983), Katz (1981)]. However, the microprocessor based implementation of real-time digital controlling is creating more problems, particularly when the computation speed and arithmetic precision are critical. In the compensator implementation, numbers which may represent control parameters or variables are realized using a sequence of finite register length binary digits which is called *finite wordlength* (FWL) for short. High precision calculations require large wordlength. However, a large wordlength arithmetic slows down the computation speed. A short wordlength representation of a control algorithm changes the controller dynamics. In particular, when the sampling rate is high as usually required in high performance digital feedback compensators, the compensator poles are

clustering near z=1 in the complex z–plane, and therefore coefficient inaccuracies can change the plant dynamics considerably. In an extreme case, such inaccuracies can lead to instability.

The finite precision performance of digital estimators and controllers has received much attention over recent years. A round–off noise analysis of certain sampled–data and direct–digital control systems have been considered in [Knowles and Edwards (1965–6), Curry (1967)]. An upper bound on the effects of quantization in digital control systems has been developed in [Johnson (1965), Slaughter (1964), Bertram (1958)]. Nowadays, digital control textbooks discuss the round–off noise analysis [Franklin and Powel (1980), Katz (1981), Åström and Wittenmark (1984)]. Fixed–point implementation of Kalman predictors has been discussed in [Sripad (1981), Scharf and Sigurdson (1984)]. The floating–point implementation of digital compensators has been examined in [Rink and Chong (1979), Van Wingerden and De Koning (1984)]. In these papers, the round–off residues have been described by considering the control parameters subject to multiplicative noise. The statistics of floating–point quantization noise can be derived by means of simulation [Phillips (1980)]. A more comprehensive analysis which includes scaling and structure issues has been investigated recently [Moroney et. al (1983), Ahmed and Belanger (1984), Sasahara et. al (1984)].

These papers have been concerned with the degradation in performance of a compensator originally designed under the assumption that the compensator will be implemented in infinite precision arithmetic. Actually, the wordlength should be directly taken into consideration when designing the compensator in the first place. In other words, the Kalman filter and the controller gains which are to be chosen for an 8–bit implementation can be very different from the corresponding gains for 16–bit implementation. This was demonstrated in a recent paper on Kalman filter design for state estimation [Williamson (1985)]. Another major point in this thesis is the demonstration of the improvement in digital control performance that can be obtained by including the finite precision nature of the implementation in the

feedback compensator design.

The linear quadratic Gaussian (LQG) problem in which the state-space representation is used has been widely investigated [Athans (1971)] and now it can be found in most digital control textbooks. The (state-feedback) compensators designed using the LQG technique are being increasingly implemented in real systems. The (scalar) LQG quadratic cost (which is also called LQG performance index) can be used as a measure of various control/estimator objectives. The performance of a classical minimum variance regulator (MVR) [Åström (1970)] which reflects the variance of the regulated output can be represented by a certain LQG performance index. The FWL Kalman filter performance considered in [Sripad (1981), Williamson (1985)] was in fact an LQG quadratic cost. The degradation in compensator performance due to FWL implementation examined in [Moroney et. al (1983), Sasahara et. al (1984)] was derived from the LQG performance index. In this thesis, we restrict the investigation to (infinite-horizon) LQG problems.

1.2 CYCLOSTATIONARITY IN THE DIGITAL REGULATION OF CONTINUOUS-TIME SYSTEMS : CONCEPTS AND CONSEQUENCES

Cyclostationary processes which are random processes with statistical properties that vary periodically with time are common in random signal analysis. In electronic communication systems where the studies of cyclostationary processes are rooted, the cyclostationary processes have been recognized since the early fifties [Deutsch (1954)]. A cyclostationary process is defined as follows.

Definition 1.1 A random process $\{x(t):-\infty<t<\infty\}$ is *wide sense cyclostationary* (WSCS) with period T_x if the mean function $\overline{x(t)}$ is periodic, and the autocorrelation function $r_x(s,t)$ is jointly periodic of period T_x; that is

$$\overline{x(t)} = \overline{x(t+T_x)} \quad \text{for all } t \tag{1.1a}$$

$$r_x(t,s) = r_x(t+T_x, s+T_x) \quad \text{for all } t \text{ and } s \tag{1.1b}$$

More generally, a *cyclostationary* process is one in wich the probability density

function (pdf) f_x and the joint pdf f_{xy} satisfy the conditions

$$f_{x(t)}(z) = f_{x(t+T)}(z) \quad \text{for all } t \text{ and } z \tag{1.2a}$$

$$f_{x(t)x(s)}(z_1,z_2) = f_{x(t+T)x(s+T)}(z_1,z_2) \tag{1.2b}$$

for all s, t, z_1 and z_2.

□□□

Cyclostationary processes are usually produced when stationary processes are subjected to some form of periodic operation such as sampling and scanning. The perodic operation is often introduced intentionally to put the signal in a format which can be manipulated easily. A specific example is the output of a receiver connected to a narrow-beam radar antenna which circularly scans a field of stationary signal sources. The statistics of the receiver output vary periodically at the rate of the revolution of the antenna [Franks (1986)]. Other examples include a television signal which is obtained by rectangular scanning of a random video field, synchronous multiplexing schemes, and synchronizing and framing techniques employed in data transmission.

The periodic operators which are frequently encountered in digital control applications are the analog-to-digital converter (ADC) and the digital-to-analog converter (DAC) systems. Specifically, consider the zero-order sample-and-hold operation [Franks (1969)] which converts the WSS input signal z(t) into an output signal y(t). The output signal y(t) is related to the input signal z(t) by

$$y(t) = \sum_{k=-\infty}^{\infty} z(kT_m)p(t-kT_m) \tag{1.3a}$$

where $k=0,\pm1,\pm2,\ldots$ and T_m is the sampling period and where the unit pulse p(t) is given by

$$p(t) = \begin{cases} 1 & \text{for } 0 \leq t < T_m \\ 0 & \text{otherwise} \end{cases} \tag{1.3b}$$

The mean of the output signal y(t) can be derived from (1.3a) and is given by

$$y(t) = \sum_{k=-\infty}^{\infty} \xi\{z(kT_m)\}p(t-kT)$$

but $\xi\{z(kT_m)\} = \overline{z(t)}$ and

$$\sum_{k=-\infty}^{\infty} p(t-kT) = 1$$

Hence

$$\overline{y(t)} = \overline{z(t)} \tag{1.4}$$

From (1.3a), the autocorrelation function $r_y(t+\delta,t)$ of the signal y(t) can be evaluated as

$$r_y(t+\delta,t) = \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \xi\{z(kT_m)z(jT_m)\}p(t+\delta-kT_m)p(t-jT_m) \tag{1.5}$$

Define the periodic indicator fucntion $q(t,\delta)$ as follows

$$q(t,\delta) = \sum_{k=-\infty}^{\infty} p(t+\delta-kT_m)p(t-kT_m) \tag{1.5b}$$

From p(t) in (1.3b), the indicator function $q(t,\delta)$ is given by

$$q(t,\delta) = \begin{cases} 1 & 0 \leq t < T_m-\delta \\ 0 & \text{otherwise} \end{cases} \tag{1.5c}$$

Let j=k+n in (1.5). The autocorrelation function $r_y(t+\delta,t)$ can be rewritten as

$$r_y(t+\delta,t) = \sum_{n=-\infty}^{\infty} r_z(nT_m)q(t,\delta+nT_m) \tag{1.6}$$

where $q(t,\delta)$ is defined in (1.5b) and $r_z(nT_m)$ is the autocorrelation function of the signal z(t). From (1.4) and (1.6) and from definition 1.1, the output signal y(t) is cyclostationary of period T_m.

The sample-and-hold operation that we considered above explains the input-output characteristics of a periodic system (A periodic system is a system which contains some form of periodic operator such as samplers or scanners). It has been shown in [Gardner (1972)] that an arbitrary (stable) periodic system whose

input is WSS will produce a WSCS output. The DAC and the ADC systems certainly belong to a class of periodic systems. We now examine an intuitively obvious fact. That is, the preservation of the property of cyclostationarity by linear time-invariant (stable) systems. Specifically, consider a linear time-invariant system which can be characterized by an impulse response function $h(t-\sigma)$. Assume the input u(t) is a WSCS of period T_c. The output y(t) is related to the input u(t) by

$$y(t) = \int_{-\infty}^{\infty} h(t-\sigma)u(\sigma)d\sigma \tag{1.7}$$

The mean of the output $y(t+T_c)$ can be derived from (1.7), we have

$$\overline{y(t+T_c)} = \xi\{y(t+T_c)\}$$

$$= \int_{-\infty}^{\infty} h(t+T_c-\sigma)\overline{u(\sigma)}d\sigma \tag{1.8}$$

Let $\sigma=\tau+T_c$, from (1.11a) we obtained

$$\overline{y(t+T_c)} = \int_{-\infty}^{\infty} h(t-\tau)\overline{u(\tau+T_c)}d\tau$$

But $\overline{u(\tau+T_c)} = \overline{u(\tau)}$ and hence

$$\overline{y(t+T_c)} = \overline{y(t)} \tag{1.9}$$

From (1.7), we can derive the autocorelation function $r_y(t+T_c,s+T_c)$, we get

$$r_y(t+T_c,s+T_c) = \xi\{y(t)y(s)\}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t+T_c-\sigma)h(s+T_c-\tau)r_u(\sigma,\tau)d\sigma d\tau \tag{1.10}$$

Let $\sigma=\gamma+T_c$ and $\tau=\nu+T_c$, from (1.10) we obtain

$$r_y(t+T_c,s+T_c) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\gamma)h(s-\nu)r_u(\gamma+T_c,\nu+T_c)d\gamma d\nu \tag{1.11}$$

But u(t) is WSCS of period T_c which implies $r_u(\gamma+T_c,\nu+Tc) = r_u(\gamma,\nu)$ and hence

$$r_y(t+T_c,s+T_c) = r_y(t,s) \tag{1.12}$$

where $r_y(t,s)$ is given by (1.10) for $T_c=0$. From (1.9) and (1.12) and from definition 1.1, the output y(t) is a WSCS of periodic T_c.

In the design of digital state-feedback compensators, there are two sub-problems to be solved. The first problem concerns the selection of feedback control law $u(kT_c)$ while the second problem deals with the reconstruction of the state vector $\hat{x}(kT_c)$. The conventional state-feedback control law design is entirely based on the knowledge of the system at the controlling instants $t_k=kT_c$ ignoring the cyclostationarity of the continuous-time response. In fact, the intersample behavior should be taken into consideration to improve the continuous-time response of digitally controlled continuous-time systems.

Synchronous (and uniform) sampling of measurement and control allows the optimal solution to the state reconstruction problem to be derived when both process disturbance and measurement noise are WSS. Synchronous measurement and control is not necessary optimal when the cyclostationary processes are considered in the design of state reconstruction.

1.3 FINITE WORDLENGTH EFFECTS IN DIGITAL CONTROL APPLICATIONS : AN OVERVIEW

If an infinite precision arithmetic were available for implementation, the linear quadratic Gaussian (LQG) problems are limited to the problem of finding the Kalman filter and controller gains which minimize the prescribed LQG performance index subject to the stability constraint. In practice, the compensators are implemented using finite wordlength (FWL) arithmetic. The effects of FWL implementation of digital compensators have been investigated in recent years by several authors. A very recent review may be found in [Hanselmann (1987)].

There are several factors which determine how the performance of digital compensators are affected by the FWL implementation. The main factors include

a. arithmetic format

b. quantization noise

c. compensator structure

d. dynamic range and scaling

We briefly review each of the factors.

1.3.1 Arithmetic Formats

Numbers which may represent either compensator parameters or variables in the digital computer must be taken in, stored, calculated and put out with finite accuracy. There are several choices of arithmetic which could be used to represent numbers. However, there are two types of arithmetic that commonly used both in digital signal processing and in digital control application; namely the *fixed-point* and the *floating-point* arithmetics. For a floating-point representation, the value of a number is represented using two numbers: the fraction or mantissa and the exponent. The values of the fraction and the exponent are both represented in fixed-point formats. The advantage of fixed-point arithmetic is that it has a faster computation speed and it requires less expensive hardware to implement. The floating-point arithmetic provides a larger dynamic range for the signals in the compensator. The arithmetic overflow which may occur due to the limited dynamic range of the fixed-point arithmetic in many cases can be prevented by introducing appropriate scaling factors to the compensator inputs, outputs, coefficients and states. Round-off noise (due to state quantization as discussed in section 1.3.2) analysis is much simpler for the fixed-point arithmetic since the round-off noise is additive [Sripad and Snyder (1977), Barnes et. al (1985)] whereas, the round-off noise of the floating-point arithmetic is multiplicative [Fettweis (1974), Rink and Chong (1979)]. In terms of the computation speed and the hardware complexity, the fixed-point arithmetic is more favorable.

The non-standard arithmetics include *block-floating-point* arithmetic [Oppenheim (1970), Williamson et. al (1985)], *logarithmic* arithmetic [Kingsbury and Rayner (1971), Lang (1984)] and *residue number* systems [Soderstran (1977), Jenkins (1979), Tan and McInnis (1982)]. Block-floating-point arithmetic is a compromise between the fixed-point and the floating-point arithmetics. In block-floating-point arithmetic, only one exponent register is used for all variables. Excluding the exponent register, block-floating-point arithmetic is similar to fixed-point arithmetic.

The use of one exponent register in block-floating-point provides a larger dynamic range (dynamic range will be defined in section 1.3.4) than could be achieved by the fixed-point arithmetic alone. The main feature of the logarithmic arithmetic is that the quantized values are logarithmically (ie. unevenly) spaced. The round-off noise of the logarithmic arithmetic is similar to the round-off noise of the floating-point arithmetic in that they are both multiplicative (and relative). The use of logarithmic arithmetic for digital control application has been investigated in [Lang (1984), Lamaire and Lang (1986)]. Residue number systems decompose the arithmetic operations such that parallel processing of sub-operations is possible in principle. This decomposition speeds up the computation time. The use of the residue number system has been investigated in [Tan and McInnis (1982)].

1.3.2 Quantization Noise Model

In the implementation of digital compensators, parameters and variables must be represented using finite number of bits. In practice, it is required to implement the least expensive compensator. One way to achieve this requirement is to select registers which can be used to represent the parameters and the variables as efficiently as possible; that is, to use as few bits as possible. However, short wordlengths change the dynamics and the performance of the compensators. The FWL effects on the dynamics and the performance of the compensators can be studied via the quantization noise model.

Basically, there are two type of quantization commonly used in digital signal processing. A (non-complex) number r is represented by the nearest number Q[r] which can be expressed using the available wordlength (determined by number of bits). This type of quantization is known as *round-off* quantization [Phillips and Nagle (1984)]. Another type of quantization is called *truncation* quantization [Phillips and Nagle (1984)]. The truncation is performed simply by chopping the least significant bits. Therefore, no extra hardware required in a truncation quantization. However, the quantization noise effects are reduced in a round-off quantization. The round-off quantization introduces no bias. Furthermore, the round-off effects are

more easily analyzed. In this thesis, the round-off quantization is used throughout. The round-off effects in digital control applications can be divided into four areas as follows.

a. coefficient (or parameter) quantization

b. state (or node) quantization

c. input (ADC) quantization

d. overflow

In this thesis, we assume all parameters and variables are appropriately scaled such that overflow can be avoided. Consequently, the overflow issue will not be considered.

Coefficient quantization : Coefficient quantization can change the dynamics of compensator considerably. Particularly when the sampling rate is (very) high, the poles of the compensators are clustering near the unit circle in the complex z-plane. A coarse quantization can lead to instability. This problem has been recognized in the synthesis of narrow-band digital filter [Gerheim (1984), Williamsom (1987a)] for which the poles are gathering near z=1 in the complex z-plane.

The noise model of the coefficient quantization is to replace a quantizer by a linear gain block followed by an injection of an additive residue term. The analysis of the residue term has been investigated using a deterministic approach [Kawamata and Higuchi (1985)] and also using a statistical approach [Knowles and Olcayto (1968)]. In the deterministic analysis, the quantized value Q[z] of a number z is modelled as

$$Q[z] = z + \Delta z \tag{1.13}$$

where the residue term Δz is (completely) determined by coefficient wordlength B_c and coefficient value z. Using the statistical approach, the residue term Δz in (1.13) is modelled as a zero-mean 'white' noise process having covariance

$$\xi\{(\Delta z)^2\} = qr(z) \tag{1.14a}$$

where q is defined by

$$q = \frac{1}{12}\, 2^{-2B_c} \tag{1.14b}$$

where B_c is the coefficient wordlength, and where r(z) is a function defined by

$$r(z) = \begin{cases} 1 & \text{if } z \text{ is non-integer} \\ 0 & \text{if } z \text{ is integer} \end{cases} \tag{1.14c}$$

It has been shown in [Kawamata and Higuchi (1985)] that the statistical analysis of *finite coefficient wordlength* (FCWL) effects in digital filters is only an approximation to the actual effects. Nevertheless, the FCWL effects are more easily analyzed using the statistical approach. This is the approach used in [Sasahara et. al (1984)].

A large coefficient wordlength slows down the computation speed of a digital computer. Therefore, a short wordlength is favorable. However, a short wordlength introduces high inaccuracies. Several discrete optimization techniques have been proposed for reducing the coefficient wordlength [Charalambous and Mitra (1972), Smith (1979), Kwan (1979)]. Another method of reducing the coefficient wordlength known as *coefficient residue correction* (CRC) is proposed in [Williamson (1985a)]. Specifically consider the following term

$$f_{ij}{}^{*}Q[x_j{}^{*}(kT_c)] \tag{1.15a}$$

where $f_{ij}{}^{*}$ is the $(i,j)^{th}$–element of the quantized coefficient f_{ij} of a coefficient matrix F and $Q[x_j{}^{*}(kT_c)]$ is the j^{th}–element of the quantized state $Q[x^{*}(kTc)]$ of the state vector $x(kT_c)$. The CRC scheme basically replaces the term in (1.15a) by

$$f_{ij}{}^{*}Q[x_j{}^{*}(kT_c)] + \varphi_{ij}Q[x_j{}^{*}(kT_c)]_H \tag{1.15b}$$

where $Q[x_j{}^{*}(kT_c)]_H$ is the twice quantized value of $Q[x_j{}^{*}(kT_c)]$ and φ_{ij} is an integer restricted to a power of two so as to maintain the wordlength consistency.

State quantization : Consider a node signal (or state) which is represented with B fractional bits (this signal is therefore always less than one in magnitude) and consider also a coefficient which is represented by B_c fractional bits. Multiplication of the state and the coefficient will produce a product with $B+B_c$ fractional bits. This product must be quantized back to B bits to maintain wordlength consistency.

This quantization process is called *state quantization.*

The quantized state $Q[x^*(k)]$ of the state vector $x^*(k)$ is given by

$$Q[x^*(kT_c)] = x^*(kT_c) + \epsilon(kT_c) \tag{1.16a}$$

where $\epsilon(kT_c)$ is the *finite state wordlength* (FSWL) residue. Under certain conditions [Sripad and Snyder (1977), Barnes et. al (1985)], the fixed-point state quantization noise $\epsilon(kT_c)$ in (1.16a) can be modelled as a zero-mean 'white' noise process having covariance

$$\xi\{\epsilon(kT_c)\epsilon'(kT_c)\} = qI_x \tag{1.16b}$$

where I_x is an $(n_x x n_x)$ identity matrix where n_x is the dimension of the state vector $x(kT_c)$ and where q is defined by

$$q = \frac{1}{12}\, 2^{-2B} \tag{1.16c}$$

where B is the state wordlength. Equation (1.16b) implies that each of the states is represented using equal fractional bits B. In general, such restriction is not necessary.

The FSWL effects on the compensator performance can be reduced by using the so called *error spectrum shaping* (ESS) [Abu-El-Haija and Peterson (1979)] technique or by using the *integer residue correction* (IRC) scheme proposed in [Williamson and Sridharan (1985b)]. Specifically consider the following term

$$f_{ij}{}^* x_j{}^*(kT_c) \tag{1.17}$$

where $f_{ij}{}^*$ is defined in (1.15a) and $x_j{}^*(k)$ is the j^{th}-element of the state vector $x^*(kT_c)$. Substitute $x_j{}^*(kT_c)$ in (1.17) by the j^{th}-element of (1.16a) results in

$$f_{ij}{}^* x_j{}^*(kT_c) = f_{ij}{}^* Q[x_j{}^*(kT_c)] + f_{ij}{}^* \epsilon_j(kT_c) \tag{1.18a}$$

The IRC scheme replaces the term $f_{ij}{}^* Q[x_j{}^*(kT_c)]$ in (1.18a) by

$$f_{ij}{}^* Q[x_j{}^*(kT_c)] + \Psi_{ij}\epsilon_j(kT_c) \tag{1.18b}$$

where $\epsilon_j(k)$ is the j^{th}-element of the residue vector $\epsilon(k)$ defined in (1.16a) and the coefficient correction Ψ_{ij} for each i and j is restricted to be an integer (often nearest to $f_{ij}{}^*$) to avoid an extra multiplication. When $f_{ij}{}^*$ is close to +1, a choice of Ψ_{ij}=-1 will greatly reduce the influence of the noise $\epsilon_j(kT_c)$ in (1.18a-b).

ADC quantization : The analog-to-digital (ADC) placed at the compensator input (or plant output) converts the analog signal $y_j^*(t)$ (the j^{th}-element of the vector $y^*(t)$) into a digital representation $Q[y_j^*(kT_c)]$ using B_{ad} bits. The quantized $Q[y_j^*(kT_c)]$ is related to the actual $y_j^*(kTc)$ by

$$Q[y_j^*(kT_c)] = y_j^*(kT_c) + \delta_j(kT_c) \tag{1.19a}$$

where the residue $\delta(kT_c)$ under certain conditions [Sripad and Snyder (1977), Barnes et. al (1985)] can be modelled as a zero-mean 'white' noise process having covariance

$$\xi\{(y_j^*(t))^2\} = \frac{1}{12}\, 2^{-2B_{ad}} \tag{1.19b}$$

where B_{ad} is the ADC wordlength. In practice, the ADC is (physically) separated from the (main) compensator. Therefore, the restriction to set B_{ad}=B is not necessary. In fact, it is common to make $B_{ad}\neq B$. In order to avoid the effects of ADC quantization on the compensator peformance, B_{ad} should be large. However, the signal y(t) contains a certain quantity of noise. Therefore, a high precision ADC may not be useful. In summary, the required ADC worlength is primarily determined by the signal-to-noise ratio (SNR) of the signal to be converted.

1.3.3 Compensator Structures

In an infinite precision arithmetic, all minimal state-space representations of the same transfer function are equivalent. Therefore, different coordinate bases do not affect the compensator performance. However, compensator structure plays an important role in the FWL compensator design. Various criterions can be used to measure the FWL performance of a particular structure. The following measures are frequently used

a. dynamic range and scaling constraints

b. round-off noise performance

c. coefficient sensitivity

d. computation speed

e. multivariable capability

The first four of the above criterions are common in digital filtering. The dynamic range and scaling constraints issues will be discussed separately in sub–section 1.3.4. In this sub–section, we briefly discuss the round–off noise performance and the coefficient sensitivity measures. The speed and the multivariable capability issues will be mentioned in the discussion of a particular structure below.

Round–off noise performance : In sub–section 1.3.2, we have discussed the round–of noise (due to state quantization) model. The variance of each round–off noise source is given by q where q is defined by (1.16c). If a digital compensator were seen as a stand alone digital filter then the variance of compensator output would be a good measure of performance. However, in digital control applications, the overall (ie. closed–loop) performance is the most important consideration. Consequently, the LQG performance index J (which will be discussed in detail in chapters 4, 5 and 6) is used to measure and to compare the round–off noise performance of different compensator structures.

Note that the comparison of round–off noise performance of different compensator structures is valid only if the compensator is properly scaled (the scaling issue will be discussed in sub–section 1.3.4). Let J_∞ and J_q be the ideal (ie. infinite precision) and the FWL performance indices. The FWL effects on the compensator performance can be written as

$$\Delta J \triangleq J_q - J_\infty \tag{1.20}$$

The structure optimization procedures developed in [Mullis and Roberts (1976), Hwang (1977), Williamson (1986)] can be carried out using ΔJ in (1.20). The investigation of round–off noise peformance of digital compensators examined in [Moroney et. al (1983), Sasahara et. al (1984)] was essentially based on the index ΔJ defined in (1.20). The FWL performance of the estimators investigated in [Sripad (1981), Williamson (1985)] was also measured using ΔJ in (1.20).

Coefficient sensitivities : In sub–section 1.3.2, we have shown that approximating the coefficients (or parameters) of a compensator with a finite number of bits will change the compensator dynamics and hence will cause a degradation in the overall

performance. For simplicity, all compensator coefficients are represented using an equal wordlength B_c. Although such restriction is not necessary in general. There are several ways to examine the effects of coefficient quantization in digital control applications. A simple way is to re-compute the phase and gain margin via frequency response method (eg. Bode plot, Nyguist plot etc.) for set of quantized coefficients. Using this approach, the coefficient worlength B_c is then selected as the shortest wordlength such that the phase and gain margin meet the design requirement. Another approach is the so called *pole-zero sensitivities* [Williamson (1986)]. This technique examines the (first order) sensitivities of the poles and zeros of a realization to variations (due to coefficients quantization) in the parameters of the realization. In digital filter design, a statistical approach based on first order sensitivities has been investigated [Crochiere (1975)]. In this technique, the shortest wordlength is estimated such that the amplitude of the filter transfer function is within a prescribed error bound. The resulting wordlength known as the *statistical wordlength* is generally non-integer. The actual wordlength is obviously selected to be the nearest integer to the statistical wordlength. The use of the statistical wordlength approach in digital control applications has been investigated in [Moroney et. al (1980)]. Using the LQG performance index, the statistical worlength technique can not be derived using the first order sensitivities because the optimal nature of the LQG design demands all the first (partial) derivatives of the index J with respect to compensator coefficients are zero. Consequently, a higher order approximation is necessary. In [Moroney et. al (1980)], a statistical wordlength formulation which was derived using the second order sensitivities is used to measure the effects of coefficient quantization. The proposed formula contains two terms. The first term represents the number of bits necessary to represent the integer portion of the coefficients and the second term provides the required number of bits for the fractional part of the coefficients. Another approach which is also based on the LQG quadratic cost is proposed in [Sasahara et. al (1984)]. This method which is an extension of the method developed in digital filter design [Kawamata and Higuchi

(1985)] is based on ΔJ defined in (1.20) but now J_q is replaced by J_c which represents the compensator performance for the quantized coefficients.

It is argued in [Sasahara et. al (1984)] that the structure which minimizes the round-off noise effects (eg. the optimum structure proposed in [Mullis and Roberts (1976), Hwang (1977), Williamson (1986)]) will also minimizes the effects of coefficients quantization. This fact has also been shown to be true in digital filter design [Jackson (1976)]. However, in digital filters [Bonzanigo (1974)], it is shown that low coefficient sensitivities does not necessarily guarantee a low round-off noise.

In an infinite precision state space realization, there are (infinitely) many structures which will give an equal input-output relation because one state space description can be transformed into another description by a similarity transformation. Specificallly, consider a (discrete) state space realization of the form

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) \tag{1.21a}$$

$$y(kT_c) = Lx(kT_c) + Du(kT_c) \tag{1.21b}$$

Another realization can be derived from the realization (1.21a-b) by using a (non-singular) similarity transformation T, results in

$$z((k+1)T_c) = \tilde{\Phi} z(kT_c) + \tilde{\Gamma} u(kT_c) \tag{1.22a}$$

$$y(kT_c) = \tilde{L} z(kT_c) + Du(kT_c) \tag{1.22b}$$

The system matrices $\{\tilde{\Phi},\tilde{\Gamma},\tilde{L}\}$ in (1.22a-b) are related to the system matrices $\{\Phi,\Gamma,L\}$ in (1.21a-b) by

$$\tilde{\Phi} = T^{-1}\Phi T \quad ; \quad \tilde{\Gamma} = T^{-1}\Gamma \quad ; \quad \tilde{L} = LT$$

The input-output characteristics of realization (1.21a-b) is equivalent to that of realization (1.22a-b) and it is given by a discrete polynomial H(z)

$$H(z) = D + L(zI-\Phi)^{-1}\Gamma \tag{1.23a}$$

$$= D + \tilde{L}(zI-\tilde{\Phi})^{-1}\tilde{\Gamma} \tag{1.23b}$$

The use of different structures for implementing a compensator with pre-computed gains has received much attention in recent years. A survey paper on digital controllers implementation [Hanselmann (1987)] contains a review section on compensator structures. Below, we briefly discuss structures which are frequently

used.

Direct form structures : Given a transfer function H(z), a state–space structure can be derived directly from the coefficients of H(z). The corresponding structure is called *direct form* (DF) structure. There are several forms of DF structures [Phillip and Nagle (1984), Oppenheim and Schafer (1975)]. The common DF structures are the controllable and the observable canonical DF structures [Kailath (1980), Kuo (1980)]. They are called canonical because they require the minimum number of additions, multiplications and delays to implement a particular realization. Therefore, in terms of structural complexity (which is measured by the number of multiplications required per output sample) DF structures are the simplest structures to implement. In terms of computation speed, DF structures posses high speed capability. However, there are drawbacks in DF structures. The first undersirable property of DF structures is that the non–zero and the non–unity coefficients can be spread over a very large dynamic range [Hwang (1975b)]. Another drawback is that DF structures produce high round–off noise and also sensitive to parameter variations due to coefficient quantization [Chan et. al (1973), Fettweis (1972), Jackson (1976)].

Parallel and cascade construction of 2^{nd}–order DF structures : One way to avoid the handicaps of (high order) DF structures is to construct several 2^{nd}–order DF structures in series (cascade) and/or parallel to form a high order structure [Jackson (1970b)]. A *cascade structure* is formed when all 2^{nd}–order DF structures are constructed in series. In the sense of round–off noise performance, optimal distribution of poles and zeros of cascade structures are permissible [Rader (1982)]. Furthermore, the round–off noise effects of each internal 2^{nd}–order DF structure can also be minimized. However, the series configuration slows down the computation speed because output (at the end block) appears after calculations in all blocks have been completed. A *parallel structure* is formed when all 2^{nd}–order DF structures are put in parallel. A multi–input multi–output compensator can be implemented using a parallel structure. A third type of the structure is a combination of parallel and cascade construction of 2^{nd}–order DF structures.

Delay replaced direct form structures : Another way to reduce the undersirable properties of (high order) DF structures is to replace each delay component Z^{-1} in a DF structure by

$$Z^{-1} = \frac{\gamma z^{-1}}{1+\delta z^{-1}} \qquad (1.24)$$

where γ and δ are certain (scalar) constants. The resulting structure is termed *delay replaced direct form* (DRDF) structure. Several types of DRDF structures which correspond to different values of γ and δ in (1.24) have been proposed in the digital signal processing [Agarwal and Burrus (1975), Szczupak and Mitra (1978), Nishimura et. al (1981), Orlandi and Martinelli (1984)]. In control applications, a delay operator which corresponds to $\gamma=\delta=1$ in (1.24) [Middleton and Goodwin (1986)] is used as a mean of improving the deterministic characteristic of digitally (finite wordlength) controlled plants. Recently, another DRDF structure which suggests to replace the k^{th} delay operator Z^{-1} in a DF structure by

$$Z^{-1} = \frac{\gamma_k z^{-1}}{1+z^{-1}} \qquad (1.25)$$

is proposed in [Williamson (1987a)]. The scalar γ_k in (1.25) for each k is selected to satisfy a certain scaling constraint (will be discussed in sub-section 1.3.4). The resulting structure is called *scaled* DRDF structure. It has been shown that the sensitivity (due to state and coefficient quantization) performance of filters implemented using scaled DRDF structures can be improved. The improvement is significant for low pass narrow band filters.

Default structure : State-space descriptions are often derived directly from a mathematical model (of a plant). In this realization, each state (or node) has a physical meaning. Compensator structures in which each state $\hat{x}_j(kT_c)$ provides an estimate of the state $x_j(k)$ of the physical model are therefore called *default structures*. The FWL performance of compensators implemented using simple (or default) structures has been investigated in [Moroney et. al (1983)]. A disadvantage of the default structures is that they may contain an excessive number of non-zero

and non–unity coefficients. Therefore, more multiplications per output sample are required than would be required by other structures.

There are many more types of structure available such as *ladder and lattice structures, wave digital filters, normal* or *coupled forms* [Oppenheim and Schafer (1975), Rabiner and Gold (1975)]. However, most of these structures are suitable only for digital filters. Another structure which is suitable for digital control applications is the *optimal* structure proposed in [Mullis and Roberts (1976), Hwang (1977), Williamson (1987a)]. This structure minimizes the round–off noise performance (and also reduces the effects of coefficient quantization). However in general, this structure suffers from the fact that it requires maximum number of multiplications per output sample (ie. the highest complexity structure). Nevertheless, in terms of round–off noise performance (and coefficient sensitivities) the optimal structure is good as a 'bench–mark' for comparing the FWL performance of different compensator structures.

1.3.4 Dynamic range and scaling constraints

In a finite precision algorithm, numbers either to represent compensator coefficients or variables are represented using finite number of bits. The *dynamic range* of a certain representation is defined to be the total numerical range of numbers that can be represented by the available number of bits. When the numerical values of certain numbers exceed the dynamic range, it is said that (numerical) *overflow* has occured. An *underflow* occurs when certain numbers can not be 'exactly' represented using the digits available. In fixed point arithmetic, overflow (and underflow) are likely to occur because of the limited dynamic range nature of the arithmetic. The occurrence of overflow can be avoided if the parameters or variables to be represented are scaled such that the dynamic range of parameters or variables matches the dynamic range of the numerical representation. Several scaling methods have been extensively discussed [Jackson (1970a), Hwang (1975a–b)].

Consider a state–space representation (1.21a–b). Assume the (discrete) transfer

function U(z) of the sequence $\{u(kT_c):-\infty<k<\infty\}$ exists. The amplitude of the j^{th} state $x_j(t)$ is then given by

$$|x_j(kT_c)| \leq \|H_j(\omega)\|_p \|U(\omega)\|_s \tag{1.26a}$$

where $p \geq 1$ and s satisfy

$$\frac{1}{p} + \frac{1}{s} = 1 \tag{1.26b}$$

and where the vector transfer function $H(\omega)$ is given by

$$H(\omega) = (zI-\Phi)^{-1}\Gamma \tag{1.26c}$$

where Φ and Γ are given in (1.21a). The L_p-norm denoted by $\|.\|_p$ in (1.26a) is defined by

$$\|H(\omega)\|_p \triangleq \{\frac{1}{\omega_m} \int_0^{\omega_m} |H(\omega)|^p d\omega\}^{1/p} \tag{1.27}$$

where $\omega_m=2\pi/T_c$, and the operations on the vector function $H(\omega)$ are done for each of its components. A *frequency domain scaling criterion* [Jackson (1970a)] is defined by

$$\|\tilde{H}(\omega)\|_p = [1\ 1\ \ldots\ 1]' \tag{1.28}$$

where $\tilde{H}(\omega)$ is the scaled version of $H(\omega)$. If $\|U(\omega)\|_s \leq \bar{u}$ for some constant $\bar{u}$ where s satisfies (1.26b) then from (1.26a) and (1.28) we have $|x(kT_c)| \leq \bar{u}[1\ 1\ \ldots\ 1]'$. From (1.26c), the criterion (1.28) can be interpreted as the unity average gain in the sense of L_p-norm (1.27) of the transfer function from input $u(kT_c)$ to each state $x_j(kT_c)$ over the frequency range $[0,\omega_m]$. It has been shown in [Hwang (1975b)] that the L_p-norm (1.27) is monotonically increasing; that is

$$\|H(\omega)\|_\infty \geq \ldots \geq \|H(\omega)\|_p \geq \|H(\omega)\|_s \geq \ldots \geq \|H(\omega)\|_1 \tag{1.29}$$

where $\infty \geq p > s \geq 1$.

There are three scaling policies that are frequently considered, namely L_∞-norm (ie. bounding the frequency spectrum), L_2-norm (ie. bounding the mean-squared values of the magnitude of the variable) and L_1-norm (ie. bounding the (absolute) amplitude of the variable) [Oppenheim and Schafer (1975)]. Note that the L_1-norm is the most pessimistic (or conservative) constraint. Another type of scaling method

which is based on the time-domain ℓ_p-norm has been described in [Hwang (1975a)]. Consider the case when p=2 in the frequency domain criterion (1.28) and let $h(kT_c)$ be the vector impulse response at the states. By Parseval's relation, the criterion (1.28) for p=2 implies

$$\sum_{k=0}^{\infty} \tilde{h}(kT_c).\tilde{h}(kT_c) = [1\ 1\ \ldots\ 1]' \tag{1.30}$$

where $\tilde{h}(kT_c)$ is the scaled version of $h(kT_c)$ and the multiplication on the left hand side of (1.30) is the dot product. The time domain counterpart of (1.26a) is given by [Hwang (1975a)]

$$|x(kT_c)| \leq \|h(kT_c)\|_p \|u(kT_c)\|_s \tag{1.31}$$

where $p \geq 1$ and s satisfy (1.26b).

A *time domain scaling criterion* which can be used as a sufficient condition of the overflow constraint is defined by

$$\|\tilde{h}(kT_c)\|_p = [1\ 1\ \ldots\ 1]' \tag{1.32a}$$

where the ℓ_p-norm denoted by $\|.\|_p$ in (1.32a) is defined by

$$\|h(kT_c)\|_p \triangleq \{\sum_{k=0}^{\infty} |h(kT_c)|^p\}^{1/p} \tag{1.32b}$$

If $\|u(kT_c)\|_q \leq \bar{u}$ where s satisfies (1.26b) then from (1.31) and (1.32b) we obtain $|x(kT_c)| \leq \bar{u}[1\ 1\ \ldots\ 1]'$. It has been shown in [Hwang (1975b)] that the ℓ_p-norm is monotonically decreasing; that is

$$\|h(kT_c)\|_\infty \leq \ldots \leq \|h(kT_c)\|_p \leq \|h(kT_c)\|_s \leq \ldots \leq \|h(kT_c)\|_1 \tag{1.33}$$

where $\infty \geq p > s \geq 1$. Therefore, in the ℓ_p-norm sense, ℓ_∞-norm which corresponds to bounding the maximum magnitude of the variable is the least conservative (or most optimistic) constraint. The ℓ_2-norm scaling is usually called stochastic scaling since ℓ_2-scaling forces equal probability of overflow in each state.

Using Parseval's theorem, it has been shown [Hwang (1975a)] that

$$\|H(\omega)\|_2 = \|h(kT_c)\|_2 \tag{1.34}$$

From relation (1.34), the ordering in (1.29) and (1.33) can be rearranged as follows

$$\|h(kT_c)\|_\infty \leq \|H(\omega)\|_1 \leq \|H(\omega)\|_2 = \|h(kT_c)\|_2 \leq \|H(\omega)\|_\infty \leq \|h(kT_c)\|_1 \tag{1.35}$$

Consequently, the case p=2 in both frequency and time domain constraints is the most convenient scaling method to be used for analysis.

1.4 RESEARCH AREAS AND ORGANIZATION OF THE THESIS

This thesis covers two areas of research. The first concerns a study of cyclostationarity in the digital regulation of continuous-time systems while the second studies methods for improving the FWL performance of digital LQG compensators. The first work is divided into three categories as follows.

a. Characterization of cyclostationary processes that may occur in digital control applications.

b. The consequences of cyclostationarity on the state reconstruction problem.

c. Improving intersample behavior of digitally controlled continuous-time systems.

The second investigation is also divided into three categories as follows.

a. Reducing the effects of round-off noise (due to state quantization) on the performance of LQG compensators.

b. Coefficient wordlength consideration in the design of LQG compensators.

c. Implementation of high order low-complexity low-sensitivity compensators.

Chapter 2 deals with the properties of cyclostationary processes which may occur in control applications. We review the procedures of discretizing a continuous-time system when using sampling and hold interfaces. The effects of pre-filter on the statistics of measurement noise is also considered in this chapter. The continuous quadratic cost function can be used to measure the performance of the controlled system. However, a digital design minimizes only the dicrete index. Therefore, the corresponding discrete cost is required. The continuous performance is computed by considering the intersample responses. We review the procedures for computing the intersample behavior developed in [de Souza and Goodwin (1984), Ackermann (1985)]. We also consider a series representation of cyclostationary processes [Gardner and Franks (1975), Ogura (1971)]. Specifically, we shall consider

the state space translational representation which can be used to characterize two classes of first and second order processes. The continuous responses resulting from a digital implementation of pulse amplitude modulated control signals are examined. Finally, in this chapter we explore the properties of periodic measurement noise.

Chapter 3 considers the consequences of cyclostationary disturbances for state prediction. First, we consider a single rate case and allow the measurement times to be non-synchronous with the control instants; that is when the plant output is measured at instants $kT_c+\delta$ where kT_c denotes the times when the control signal is updated and δ is a certain delay factor. The problem is to optimize the choice of δ such that the mean square prediction error is minimized. The problem can be reformulated into a standard prediction problem. The multirate case is also considered. Multirate sampling is one approach for alleviating the problem of finite computing capabilities of 'on-board' digital computers. Typically, high frequency sampling is used for plant output measurement (ie. controller input) while low frequency sampling is used for control (ie. plant input) [Glasson (1983), Broussard and Glasson (1980)]. Specifically, we consider $T_c=NT_m$ for some integer N where T_c and T_m are respectively the control and the measurement sampling periods, and the state reconstruction is based on N measurements. The optimal prediction problem concerns the selection of a group of N measurements. As with the non-synchronous prediction, the multirate estimation problem can be reformulated into a conventional prediction problem.

Chapter 4 concerns with the discrete linear quadratic regulation (LQG) of continuous-time systems. The ideal (ie. infinite precision) LQG regulator design is briefly reviewed in this chapter. We consider both finite and infinite horizon LQG problems. Specifically, we consider the separation principle [Kwakernaak and Sivan (1972)] which allows the controller and the Kalman filter problems to be solved separately. The necessary and sufficient conditions for the existence of stabilizing solutions are also considered. A minimax quadratic regulator which minimizes the worst (ie. maximum) intersample behavior is examined. We investigate the use of a

minimax output variance regulator which is a special case of a minimax quadratic regulator. It is shown that standard LQG techniques can be used to tackle the minimax quadratic regulation problem. Other methods of output regulation are reviewed. By means of examples, the performance of the proposed minimax regulator is compared to other methods for output regulation.

Chapter 5 concerns the design of finite wordlength linear quadratic regulators. As hinted earlier, the majority of results [Rink and Chong (1979), Van Wingerden and De Koning (1984), Moroney et. al (1983), Sasahara et. al (1984)] in control applications have been concerned with the degradation in performance of a compensator originally designed under the assumption that the compensator will be implemented in infinite precision arithmetic. In this chapter we consider the fixed-point FWL design which directly takes the wordlength into consideration. We shall see that the optimal FWL design involves the selection of the optimal Kalman filter and controller gains and choosing the optimal structure subject to a certain ℓ_2-scaling constraint. In this design, the integer residue correction [Williamson and Sridharan (1985b)] or the sub-optimal error spectrum shaping [Abu-El-Haija and Peterson (1982)] scheme is incorporated. The finite coefficient wordlength consideration in the design of LQG regulators is also considered in this chapter.

Chapter 6 concerns the implementation of a low-complexity low-sensitivity compensator. The optimum structure mentioned earlier suffers from the fact that it has the largest number of multiplications per output sample (ie. high complexity structure) while the direct from structure (ie. the least complex structure) produces high round-off noise. The delay replaced delay form structure (DRDF) [Agarwal and Burrus (1975), Szczupak and Mitra (1978), Nishimura et. al (1981), Orlandi and Mortinelly (1984), Williamson (1987)] is a good candidate. First, we consider a single-input single-output (SISO) DRDF implementation. In [Williamson (1987)], it is shown that the scaled DRDF structure can be derived directly from the given transfer function (hence the direct form structure). We extent this result such that the DRDF structure can be obtained directly from the given default structure. This

procedure is more appropriate for control applications. Both the controllable and observable DRDF structures are considered. The procedure can be generalized for the multi-input multi-output (MIMO) system. In this chapter we develop a procedure which allows the MIMO-DRDF structure to be derived directly from the given default structure. For both the SISO and the MIMO cases, we investigate the FWL performance of compensators implemented using the scaled DRDF structure.

Finally, in chapter 7 we present the summary, conclusions and directions for future research.

1.5 ORIGINAL IDEAS PROPOSED IN THE THESIS

State space translational representation of wide-sense cyclostationary processes as developed in chapter 2.

First order approximation of harmonic series representation of wide sense cyclostationary processes as described in chapter 2.

Characterization of a wide sense cyclostationary measurement noise as derived in chapter 2.

The discrete minimax quadratic regulation (MMQR) of continuous-time systems which directly takes the intersample behavior into account as developed in chapter 3.

The discrete minimax output variance regulation (MMVR) of continuous-time systems as described in chapter 3.

Non-synchronous control and output sampling to achieve optimal state reconstruction in the presence of cyclostationary disturbances as described in chapter 4.

The development of multirate optimal state prediction as described in chapter 4 when the process disturbance and/or the measurement noise are cyclostationary.

The development of algorithms for achieving the optimum finite wordlength linear quadratic Gaussian (FWL-LQG) regulator as discussed in chapter 5.

An extension of the procedures for deriving the delay replaced direct form (DRDF) structure as described in chapter 6.

The procedures for deriving the multi-input multi-output delay replaced direct form (DRDF) structure from the given default structure as developed in chapter 6.

CHAPTER 2

CYCLOSTATIONARY PROCESSES

2.1 INTRODUCTION

For the control of continuous-time systems it is generally more practical and advantageous to use digital control especially if the desired control law is to be adaptive. From an analytical point of view, the simplest solution is to generate the control law as a pulse amplitude modulated (PAM) signal with constant period. At the controlling instants, a linear time invariant continuous-time system under the influence of continuous *wide sense stationary* (WSS) [Papoulis (1965)] 'white noise' disturbances has an equivalent linear time invariant description in which the corresponding discrete disturbances are also WSS. The usual approach is to base the design of the discrete control sequence on this system description. Often (as in the case of linear regulation) the resulting control sequence is (asymptotically) WSS. However as it is shown in [de Souza and Goodwin 1984], the continuous control signal is not stationary but will exhibit periodic statistical properties, and as a consequence so will the continuous system state and plant output. This special case of nonstationary processes is termed *cyclostationary* [Bennet (1958)], *periodic non-stationary* [Ogura (1971)] or *periodically stationary* [Papoulis (1965), Franks (1969)].

By definition [Franks (1969)], a random process $\{x(t)\}$ for all $t\epsilon(-\infty,\infty)$ is *cyclostationary* with period T_x if

a. the probability density function (pdf) of $x(t)$ is identical to the pdf of $x(t+T_x)$ for all t.

b. the joint pdf of $x(t)$ and $x(s)$ is identical to the joint pdf pdf of $x(t+T_x)$ and $x(s+T_x)$ for all t and s.

The performance measures that we consider throughout this thesis only involve second order statistics. Therefore, a definition which is based on lower order statistics is sufficient for our purposes. A random process $\{x(t):-\infty<t<t\}$ is said to be *wide sense cyclostationary* (WSCS) with period T_x if the mean function $\overline{x(t)}$ is periodic, and the autocorrelation function $r_x(t,s)$ is jointly periodic of period T_x; that is

$$\overline{x(t)} = \overline{x(t+T_x)} \qquad \text{for all } t \tag{2.1}$$

$$r_x(t,s) = r_x(s+T_x, t+T_x) \qquad \text{for all } t \text{ and } s \tag{2.2}$$

In electronic communication systems, cyclostationary processes have been recognized since the fifties [Deutsch (1954), Bennet (1958)]. Cyclostationary processes often occur in the systems in which the stationary processes are subjected to some form of periodic operation such as sampling and scanning. Such periodic operation is often introduced intentionally to put the signal in a format which can be manipulated easily. Periodic operations which are commonly used in communications include amplitude modulation (AM), frequency shift keying (FSK), pulse amplitude modulation (PAM), time division multiplexing (TDM). Cyclostationary processes are frequently treated as if they were (purely) stationary. This can be done by averaging the statistical parameters such as mean, covariance etc. over one period. It has been shown in [Gardner (1972)] that a significant improvement in performance of certain receiver-filters can be obtained by recognizing that the input (ie. the received signal) is actually cyclostationary.

In this chapter, we investigate the properties of cyclostationary processes which may occur in control applications. The properties of the discrete equivalent representation of continuous-time linear time-invariant systems when using a zero-order hold digital-to-analog converter (DAC) and a periodic sampling analog-to-digital converter (ADC) are reviewed in section 2.2. In this section, we also review the effect of pre-filter placed before the ADC on the statistics of the measurement noise. In section 2.3, we consider the representation of cyclostationary processes. We develop the translational representation of a WSCS process which is

then used to characterize two classes of first and second order processes. We also consider the harmonic series representation of a WSCS process in this section. The relevance of cyclostationary processes in stochastic control are explored in section 2.4. We examine the continuous time response resulting from the digital implementation of pulse amplitude modulated control signals. We shall see that the intersample variance can vary significantly from that observed at the controlling instants. By means of some examples, we show that increasing the control sampling rate (in an attempt to reduce the variance at the controlling instants) in fact increases the maximum intersample variance especially for lightly damped open-loop systems. Finally, we investigate the statistical properties of cyclostationary measurement noises.

2.2 DISCRETIZATION OF CONTINUOUS-TIME SYSTEMS

In this section we briefly review the method for discretizing continuous-time time-invariant systems assuming a zero-order hold is used in the digital to analog converter and a periodic sampler is used to transform the continuous signal into a discrete signal as shown in Fig.2.1. The continuous-time system that we consider is described by

$$\dot{x}(t) = Ax(t) + Bu(t) + \omega(t) \tag{2.3a}$$

$$y(t) = Cx(t) + \eta(t) \tag{2.3b}$$

where the dimensions of the state, the input and the output are respectively given by

$$x(t) \epsilon R^{n_x} \; ; \quad u(t) \epsilon R^{n_u} \; ; \quad y(t) \epsilon R^{n_y}$$

Note that the continuous-time model (2.3a-b) contains no delay. The dimension of system matrices A, B and C are respectively $(n_x x n_x)$, $(n_x x n_u)$ and $(n_y x n_x)$. The processes $\{\omega(t)\}$ and $\{\eta(t)\}$ for all $t \epsilon (-\infty, \infty)$ are assumed to be zero-mean independent WSS processes with respective covariance functions

$$\xi\left\{\begin{bmatrix}\omega(t)\\ \eta(t)\end{bmatrix}[\omega'(t)\ \eta'(t)]\right\} = \begin{bmatrix} r_{\omega}(t,t) & r_{\omega\eta}(t,t) \\ r'_{\omega\eta}(t,t) & r_{\eta}(t,t)\end{bmatrix}$$

$$= \begin{bmatrix}\Omega_c & \Psi_c \\ \Psi'_c & \Lambda_c\end{bmatrix} \qquad (2.4a)$$

For simplicity, we assume the processes $\{\omega(t)\}$ and $\{\eta(t)\}$ are uncorrelated. Consequently, in (2.4a)

$$r_{\omega\eta}(t,t) = \psi_c = 0 \qquad (2.4b)$$

The control signal u(t) is generated by a zero-order sample-and-hold having discrete input sequence $\{u(kT_c)\}$ where T_c is the sampling period; that is

$$u(t) = \sum_{k=-\infty}^{\infty} u(kT_c)p(t-kT_c) \qquad (2.5a)$$

where p(t) is a unit pulse described by

$$p(t) = \begin{cases} 1 \text{ for } t\in[0,T_c) \\ \\ 0 \text{ otherwise}\end{cases} \qquad (2.5b)$$

The control sequence $\{u(kT_c)\}$ is assumed to be selected to minimize a certain quadratic performance index (the detail of this design will be presented in chapter 4). The resulting stabilizing control law (which provides a stable closed-loop system) is governed by

$$u(kT_c) = - Gx(kT_c) \qquad (2.6)$$

The equivalent discrete-time system of the system (2.3) for a sampling period T_c seconds can be written as follows [Kwakernaak and Sivan (1972), Sage (1974)].

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \qquad (2.7a)$$

$$y(kT_c) = Lx(kT_c) + \eta(kT_c) \qquad (2.7b)$$

where the discrete system matrices Φ, Γ and L can be computed from the continuous system matrices A, B and C as follows

$$\Phi((k+1)T_c, kT_c) \triangleq \Phi = e^{A((k+1)T_c - kT_c)}$$

$$= e^{AT_c} \tag{2.8a}$$

$$\Gamma((k+1)T_c, kT_c) \triangleq \Gamma = \int_{kT_c}^{(k+1)T_c} e^{A((k+1)T_c-\sigma)} B d\sigma$$

$$= \int_{0}^{T_c} e^{A(T_c-\sigma)} B d\sigma \tag{2.8b}$$

$$L = C \tag{2.8c}$$

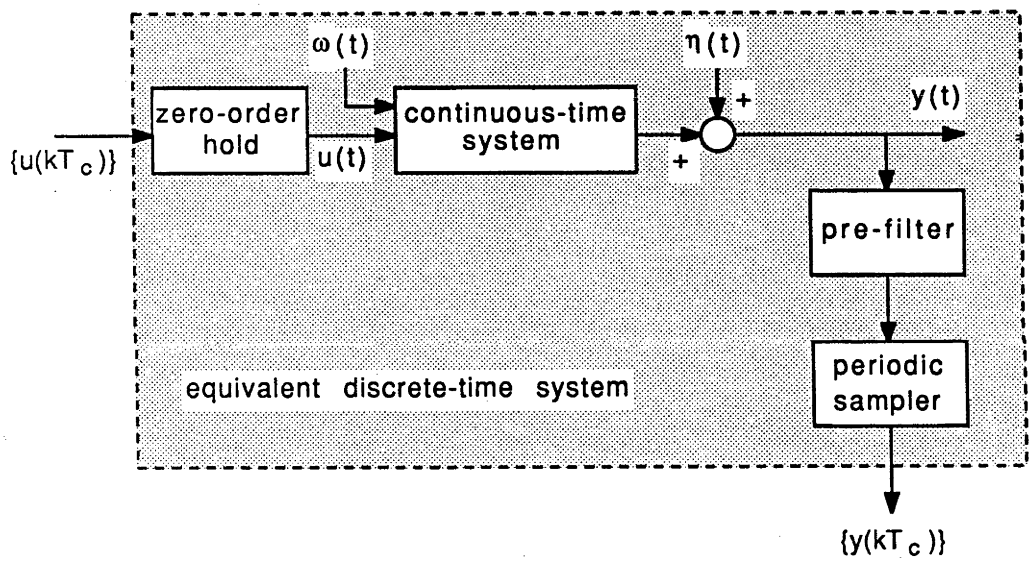


Fig.2.1 Discrete equivalent realization of a continuous-time system.

The discrete version of the disturbances $\omega(t)$ and $\eta(t)$ are given by

$$\omega(kT_c) = \int_{kT_c}^{(k+1)T_c} e^{A(T_c-\sigma)} \omega(\sigma) d\sigma$$

$$= \int_{0}^{T_c} e^{A(T_c-\sigma)} \omega(\sigma) d\sigma \tag{2.9a}$$

$$\eta(kT_c) = \eta(t)\Big|_{t=kT_c} \tag{2.9b}$$

The covariance of the discrete process $\{\omega(kT_c):-\infty<k<\infty\}$ is given by [Kwakernaak and Sivan (1972)]

$$\Omega \triangleq \xi\{\omega(kT_c)\omega'(kT_c)\}$$

$$= \xi\{\int_0^{T_c}\int_0^{T_c} e^{A(T_c-\sigma)}\omega(\sigma)\omega'(\tau)\, e^{A'(T_c-\tau)}d\sigma d\tau\}$$

$$= \int_0^{T_c} e^{A(T_c-\sigma)}\Omega_c\, e^{A'(T_c-\sigma)}d\sigma \qquad (2.9c)$$

where Ω_c is defined by (2.4a). The effects of sampling on the statistics of the measurement noise $\eta(t)$ in (2.9b) is discussed later in the section. The cross-covariance of the discrete processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$ is given by

$$\xi\{\omega(kT_c)\eta'(kT_c)\} = \xi\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{A(T_c-\sigma)}\omega(\sigma)\eta'(\tau)d\sigma d\tau\}$$

but $r_{\omega\eta}(t,t)=0$, and hence

$$\xi\{\omega(kT_c)\eta'(kTc)\} = 0$$

From (2.9a-b), it can be seen that the discretization procedure produces the zero-mean processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$.

The poles and the zeros of the discrete-time system (2.7) are related to the poles and the zeros of the continuous-time system (2.3). The relationship between the poles $\{\lambda_i(\Phi)\}$ of the discrete-time model and the poles $\{\lambda_i(A)\}$ of the continuous-time plant can be seen from (2.8a); that is [Gantmacher (1960)]

$$\lambda_i(\Phi) = e^{\lambda_i(A)T_c} \qquad (2.10)$$

for $i\epsilon[1,n_x]$ where $\lambda_i(X)$ denotes the i^{th}-eigenvalue of a matrix X. A continuous time plant is stable if all eigenvalues $\{\lambda_i(A)\}$ are located in the left-hand side of the s-plane (ie. non-positive real value as shown by the shaded-area in Fig.2.2a). From relation (2.10), the corresponding discrete-time model is stable if all eigenvalues $\{\lambda_i(\Phi)\}$ are located inside the unit-circle in the z-plane (as shown by the shaded area in Fig.2.2b).

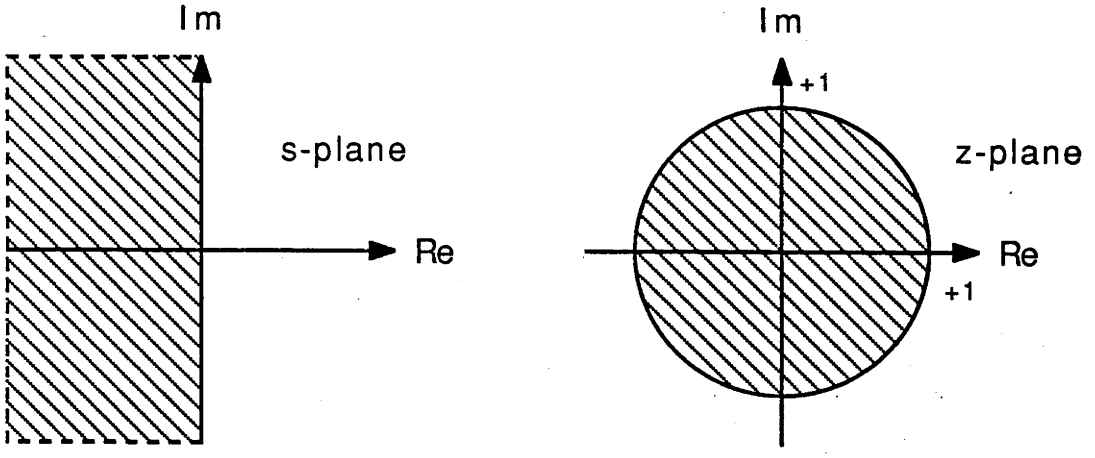


Fig.2.2 The stability regions of continuous-time and the corresponding discrete-time models.

Unlike the poles, the zeros of the discrete-time plant can not be obtained easily from the zeros of the continuous-time plant. Stable continuous-time poles (ie. poles in the left hand side (LHS) of the imaginary axis in the s-plane) are preserved in that the corresponding discrete-time poles are located inside the unit circle in the z-plane. But, minimum-phase continuous-time zeros (ie. zeros in the LHS in the s-plane) may be mapped into non-minimum phase discrete-time zeros (ie. discrete zeros outside the unit circle in the z-plane) [Åström et. al (1980)].

Assume the continuous-time representation $\{A,B,C\}$ in (2.3a-b) is controllable and observable (ie. minimal). For an arbitrary sampling period T_c, the minimality of the triplet $\{A,B,C\}$ does not imply the representation $\{\Phi,\Gamma,L\}$ in (2.7a-b) is minimal. The following lemma establishes the conditions for minimality.

Lemma 2.1 [Chen (1984)] Consider the linear continuous-time time-invariant system (2.3a-b) which is minimal. A sufficient condition for the corresponding discrete equivalent realization (2.7a-b) to be minimal is that

$$\mathrm{Im}[\lambda_i(A) - \lambda_j(A)] \neq \frac{2\pi\alpha}{T_c}$$

for $\alpha=\pm1,\pm2,\ldots$ whenever

$$\mathrm{Re}[\lambda_i(A) - \lambda_j(A)] = 0$$

where $\mathrm{Im}[\lambda]$ and $\mathrm{Re}[\lambda]$ respectively denote the imaginary and the real part of a complex number λ. For a single-input and single-output plant (ie. $n_u=n_y=1$), the above sufficient condition is also necessary.

□□□

At the controlling instants $t=kT_c$, the characteristics of the continuous system (2.3a-b) can be evaluated using the equivalent discrete-time description (2.7a-b) and (2.8a-c). But, since the system is continuous in time the intersample characteristic of the system must also be evaluated; that is, the performance must be determined at times δ for $kT_c \leq \delta < (k+1)T_c$. A method for evaluating the intersample characteristics is proposed in [de Souza and Goodwin (1984), Ackermann (1985)]. In this method, the intersample characteristics is evaluated at instants $t=n\Delta$ where $n=0,1,2,\ldots,N$ and the factor Δ is chosen such that $T_c=N\Delta$. Note that the intersample characteristics is evaluated after the characteristics at the controlling instant has been evaluated. At time $t=\Delta$ for $0<\Delta<T_c$ the equivalent discrete time system of (2.3a-b) is described by

$$x(kT_c+\Delta) = \Phi_\Delta x(kT_c) + \Gamma_\Delta u(kT_c) + \omega_\Delta(kT_c) \tag{2.11a}$$

$$y(kT_c+\Delta) = Lx(kT_c+\Delta) + \eta_\Delta(kT_c+\Delta) \tag{2.11b}$$

where

$$\Phi_\Delta = e^{A\Delta} \tag{2.12a}$$

$$\Gamma_\Delta = \int_0^\Delta e^{A(\Delta-\sigma)}d\sigma \tag{2.12b}$$

and where the discrete-time disturbances $\{\omega_\Delta(kT_c)\}$ and $\{\eta_\Delta(kT_c)\}$ are given by

$$\omega_\Delta(kT_c) = \int_0^\Delta e^{A(\Delta-\sigma)}d\sigma \tag{2.13a}$$

$$\eta_\Delta(kT_c) = \eta(t)\Big|_{t=\Delta} \tag{2.13b}$$

The covariance of the sequence $\{\omega_\Delta(kT_c)\}$ can be computed in a similar way as for

computing the covariance of the sequence $\{\omega(kT_c)\}$; that is

$$\Omega_\Delta = \int_0^\Delta e^{A(\Delta-\sigma)}\ \Omega_c\ A'(\Delta-\sigma)\ d\sigma \tag{2.14}$$

When $T_c=N\Delta$, the discrete system description (2.7a-b) is related to the system (2.11a-b) by

$$\Phi = \Phi_\Delta^N$$

$$\Gamma = \Gamma_\Delta + \Phi_\Delta\Gamma_\Delta + \ldots + \Phi_\Delta^{N-1}\Gamma_\Delta$$

$$\omega(kT_c) = \omega_{(N-1)\Delta}(kT_c) + \Phi_\Delta\omega_{(N-2)\Delta}(kT_c) + \ldots + \Phi_\Delta^{N-1}\omega_\Delta(kT_c)$$

The intersample characteristics of the system (2.3a-b) can be computed using the discrete descriptions (2.7a-b) and (2.11a-b). The result is stated in the following lemma.

Lemma 2.2 Consider the minimal continuous-time system (2.3a-b) where the signal u(t) is given by (2.5a-b) and the stabilizing control law $u(kT_c)$ is given by (2.6). Consider as well the equivalent discrete representation (2.7a-b) and (2.11a-b). Then, at any intersample instant $t=n\Delta$, $n=0,1,\ldots,N$ where $\Delta=T_c/N$, the covariance of the system can be described by

$$X(n\Delta) = \bar{\Phi}_{n\Delta}X(N\Delta)\bar{\Phi}'_{n\Delta} + \Omega_{n\Delta} \tag{2.15a}$$

$$Y(n\Delta) = LX(n\Delta)L' + \Lambda_{n\Delta} \tag{2.15b}$$

where

$$X(n\Delta) \triangleq \xi\{x(kT_c+n\Delta)x'(kT_c+n\Delta)\}$$

$$Y(n\Delta) \triangleq \xi\{y(kT_c+n\Delta)y'(kT_c+n\Delta)\}$$

$$\bar{\Phi}_{n\Delta} = \Phi_{n\Delta} - \Gamma_{n\Delta}G \tag{2.15c}$$

and where the matrices $\Phi_{n\Delta}$, $\Gamma_{n\Delta}$ and $\Omega_{n\Delta}$ are given by

$$\Phi_{n\Delta} = \Phi_\Delta^n \tag{2.16a}$$

$$\Gamma_{n\Delta} = \sum_{j=0}^{n-1} (\Phi_\Delta)^j\Gamma_\Delta \tag{2.16b}$$

$$\Omega_{n\Delta} = \sum_{j=0}^{n-1} (\Phi_\Delta)^j \Omega_c (\Phi_\Delta')^j \qquad (2.16c)$$

and $\Lambda_{n\Delta}$ is the covariance of the measurement noise at $t=n\Delta$. The covariance $X(N\Delta)$ is given by (2.15a) for n=N.

Proof : A simple extension of the discrete system description (2.11a) results in

$$x(kT_c+n\Delta) = \Phi_{n\Delta} x(kT_c) + \Gamma_{n\Delta} u(kT_c) + \omega_{n\Delta}(kT_c) \qquad (2.17a)$$

$$y(kT_c+n\Delta) = Lx(kT_c+n\Delta) + \eta_\Delta(kT_c+n\Delta) \qquad (2.17b)$$

where $\Phi_{n\Delta}$ and $\Gamma_{n\Delta}$ are given by (2.16a) and (2.16b) respectively. The discrete sequence $\{\omega_{n\Delta}(kT_c)\}$ in (2.17a) is defined by

$$\omega_{n\Delta}(kT_c) = \omega_{(n-1)\Delta}(kT_c) + \Phi_\Delta \omega_{(n-2)\Delta}(kT_c) + \ldots + \Phi_\Delta^{n-1} \omega_\Delta(kT_c) \qquad (2.18)$$

The covariance of the process $\{\omega_{n\Delta}(kT_c)\}$ is given by $\Omega_{n\Delta}$ in (2.16c). Substitution of $u(kT_c)$ in (2.17a) using (2.6b) yields the closed loop system

$$x(kT_c + n\Delta) = \bar{\Phi}_{n\Delta} x(kT_c) + \omega_{n\Delta}(kT_c) \qquad (2.19)$$

where $\bar{\Phi}_{n\Delta}$ is given by (2.15c). Evaluation of the covariance of the closed loop system (2.19) and (2.17b) gives the covariance equations (2.15a-b).

□□□

Notice that in Fig.2.1 the continuous output y(t) is filtered before it is sampled. The analog pre-filter which is in general a low-pass filter is useful for avoiding the aliasing effect. The need for such pre-filter is known in both the digital filtering area [Rabiner and Gold (1975), Oppenheim and Schafer (1975)] and in digital control applications [Åström and Wittenmark (1984), Franklin and Powell (1981)]. A conservative design procedure is to select the bandwidth of the filter to be sufficiently wide compare to the system bandwidth. Thus, ideally the pre-filter passes the Cx(t) term in (2.3b) unchanged while filtering the broadband measurement noise $\eta(t)$. To see the effects of the pre-filter, consider the first order low-pass filter governed by

$$\dot{z}(t) = -az(t) + a\eta(t) \qquad (2.20)$$

Note that the input to the filter (2.20) is $\eta(t)$ rather than $cx(t) + \eta(t)$ since it is assumed that the breakpoint of the filter (=a) is chosen such that the filter will pass

the term cx(t) unchanged. At the sampling instants $t=kT_c$ the filter response can be represented by

$$z((k+1)T_c) = e^{-aT_c}z(kT_c) + \int_0^{T_c} e^{-a(T_c-\sigma)}a\eta(\sigma)d\sigma \qquad (2.21)$$

The variance of the filter output can be derived by taking the expected value of the square of (2.21) which yields

$$\xi\{z^2(kT_c)\} = \frac{a}{2}\Lambda_c \qquad (2.22)$$

where Λ_c is the variance of the noise $\eta(t)$. In deriving the variance of $\{z(kT_c)\}$ in (2.22) the stationarity of the process $\eta(t)$ has been used to deduce that in steady state the variance of $\{z(kT_c)\}$ is constant for all k. If the bandwidth of the filter is chosen to be twice the system bandwidth (which implies $a=T_c/2$) then the variance of the filter output is inversely proportional to the sampling period T_c; that is

$$\xi\{z^2(kT_c)\} = \frac{\Lambda_c}{T_c} \qquad (2.23)$$

In general, the dynamics of the analog pre-filter should be taken into consideration in the design of digital compensators. One way to do this is to include the dynamics of the filter in the representation of the continuous-time model (2.3a-b). The resulting augmented system would have (n_x+n_f) number of states where n_f is the dimension of the pre-filter state. However, in the initial stages of design, the dynamics of the anti-aliasing filter is usually ignored. (The sensors and actuators dynamics are also ignored). Therefore, the term Cx(t) in (2.3b) is assumed to be unaffected by the pre-filter. The output y(t) in (2.3b) is approximated by

$$y(t) = Cx(t) + \tilde{\eta}(t) \qquad (2.24a)$$

where the filtered noise $\tilde{\eta}(t)$ is defined by

$$\tilde{\eta}(t) = \int_{-\infty}^{\infty} h_f(t-\sigma)\eta(\sigma)d\sigma \qquad (2.24b)$$

where $h_f(t)$ is the impulse response of the anti-aliasing filter. From (2.24a), the covariance of the discrete noise $\tilde{\eta}(kT_c)$ (an approximation of $\eta(kT_c)$ in (2.7b)) is

given by

$$\xi\{\tilde{\eta}(kT_c)\tilde{\eta}'(kT_c)\} = \int_{-\infty}^{\infty} h_f(kT_c-\sigma)\Lambda_c h_f'(kT_c-\sigma)d\sigma \qquad (2.24c)$$

As a further idealization, assume that the anti-aliasing filter is an ideal low-pass filter defined by

$$h_f(t) = \frac{1}{\pi t}\sin(2\pi f_o t) \qquad (2.24d)$$

for all t where f_o is the cut-off frequency. Note that in the frequency domain an ideal low-pass filter is characterized by the transfer function $H_f(2\pi jf_o)$ where

$$H_f(2\pi jf) = \begin{cases} 1 & \text{for } |f| \leq f_o \\ 0 & \text{otherwise} \end{cases}$$

Substitute $h_f(t)$ in (2.24c) by using (2.24d) yields

$$\xi\{\tilde{\eta}(kT_c)\tilde{\eta}'(kT_c)\} = \int_{-\infty}^{\infty} \frac{1}{\pi^2\sigma^2}\sin^2(2\pi f_o\sigma)\Lambda_c d\sigma$$

$$= 2f_o\Lambda_c \qquad (2.25a)$$

Therefore, if the cut-off frequency of the anti-aliasing filter is chosen to be twice the system bandwidth (which implies $f_o=1/(2T_c)$), the covariance of the discrete noise $\tilde{\eta}(kT_c)$ is given by

$$\xi\{\tilde{\eta}(kT_c)\tilde{\eta}'kT_c)\} = \frac{\Lambda_c}{T_c} \qquad (2.25b)$$

Throughout this thesis, we ignore the dynamics of the pre-filter and consider only the effect of pre-filter on the measurement noise $\eta(t)$. Consequently, the covariance of the discrete noise $\eta(kT_c)$ defined in (2.7b) is approximated by

$$\xi\{\eta(kT_c)\eta'(kT_c)\} = \frac{\Lambda_c}{T_c} \qquad (2.25c)$$

2.3 REPRESENTATIONS OF CYCLOSTATIONARY PROCESSES

The analysis and/or synthesis of systems involving random processes are greatly expedited by the use of appropriate representations of such processes. A random process is frequently described using an integral representation or a series representation [Franks (1969)]. In this section, the series representation is used to represent cyclostationary processes. Using the series representation approach, a cyclostationary process can be represented in two ways. The first representation is called *translational series representation* [Gardner and Franks (1975)] while the second representation is known as *harmonic series representation* [Ogura (1971), Gardner and Franks (1975)]. We first develop the relevant translational series representation and then continue with the development of the harmonic series representation.

Definition 2.1 [Gardner and Franks (1975)] A cyclostationary process $\tilde{\omega}(t)$ with period T_ω is said to have a *translational series representation* of order ℓ if there exists a complete orthonormal set of ℓ deterministic basis functions $\{\varphi_p:1\leq p\leq \ell\}$ and a set of ℓ jointly WSS random sequences $\{\{\alpha_{np}\}:1\leq p\leq \ell,-\infty<n<\infty\}$ such that

$$\xi\left\{\left|\tilde{\omega}(t) - \sum_{n=-\infty}^{\infty}\sum_{p=1}^{\ell}\alpha_{np}\varphi_p(t-nT_\omega)\right|^2\right\} = 0 \qquad (2.26)$$

for all $-\infty<t<\infty$.

□□□

In the context of this thesis, we consider a state space translational representation of $\omega(t)$ of order ℓ as defined by

$$\dot{z}(t) = Dz(t) + \sum_{k=-\infty}^{\infty} e(kT_\omega)p(t-kT_\omega) \qquad (2.27a)$$

$$\omega(t) = Hz(t) \qquad (2.27b)$$

where $z(t)\epsilon R^\ell$ and $\omega(t)\epsilon R^1$ and where D is strictly stable (ie. $\lambda_i(D)<0$ for all $1\leq i\leq \ell$) and the sequence $\{e(kT_\omega)\}$ is a zero-mean vector WSS process with covariance E. The unit pulse p(t) in (2.27a) is defined by (2.5b). The process $\omega(t)$

in (2.27b) is the output of a linear time-invariant plant driven by a zero-mean random amplitude PAM signal which is cyclostationary with period T_ω. Thus, the output $\omega(t)$ is also cyclostationary with period T_ω. At the time instants $t=kT_\omega$ for all integers k, the system (2.27a) can be represented as

$$z((k+1)T_\omega) = \Phi(T_\omega)z(kT_\omega) + \int_0^{T_\omega} \Phi(T_\omega-\sigma)e(kT_\omega)d\sigma \tag{2.28a}$$

$$\omega(kT_\omega) = Hz(kT_\omega) \tag{2.28b}$$

where

$$\Phi(\tau) = e^{D\tau} \tag{2.28c}$$

The covariance matrix $\Omega(T_\omega)$ of the process $\omega(t)$ at $t=kT_\omega$ can be derived from (2.28a) and (2.28b). The result is

$$\Psi(T_\omega) = \Phi(T_\omega)\Psi(T_\omega)\Phi'(T_\omega) + \Theta(T_\omega) \tag{2.29a}$$

$$\Omega(T_\omega) = H\Psi(T_\omega)H' \tag{2.29b}$$

where

$$\Psi(T_\omega) \triangleq \xi\{z(kT_\omega)z'(kT_\omega)\}$$

$$\Omega(T_\omega) \triangleq \xi\{\omega(kT_\omega)\omega'(kT_\omega)\}$$

and where

$$\Theta(T_\omega) = \int_0^{T_\omega} \Phi(T_\omega-\sigma)E\Phi'(T_\omega-\sigma)d\sigma \tag{2.30}$$

where

$$E \triangleq \xi\{e(kT_\omega)e'(kT_\omega)\}$$

By Lyapunov's theorem [Anderson and Moore (1979)], the steady state solution of the discrete Lyapunov equation (2.29a) $\Psi(T_\omega)$ exists, is unique and is positive definite provided the pair {F,D} is completely reachable where $\Theta(T_\omega)=DD'$ and $|\lambda_i(\Phi(T_\omega))|<1$ (ie. $\Phi(T_\omega)$ is strictly stable). The covariance matrix of the process $\{\omega(t)\}$ within a period T_ω can be evaluated by using the representation of system (2.27a) for all instants $kT_\omega \leq \delta \leq (k+1)T_\omega$; that is

$$z(kT_\omega+\delta) = \Phi(\delta)z(kT_\omega) + \int_0^{\delta} \Phi(\delta-\sigma)e(kT_\omega)d\sigma \tag{2.31a}$$

$$\omega(kT_\omega+\delta) = Hz(kT_\omega+\delta) \tag{2.31b}$$

where $\Phi(\delta)$ is given by (2.28c) for $\tau=\delta$. The covariance matrix of $\omega(t)$ for $t=\delta$ is then given by

$$\Psi(\delta) = \Phi(\delta)\Psi(T_\omega)\Phi'(\delta) + \Theta(\delta) \tag{2.32a}$$

$$\Omega(\delta) = H\Psi(\delta)H' \tag{2.32b}$$

where the covariance matrix $\Psi(T_\omega)$ is given by (2.29a) and the matrix $\Theta(\delta)$ is given by (2.30) for $T_\omega=\delta$. The characteristics of $\omega(t)$ as described by the covariance equation (2.32a) is illustrated in the following examples.

Example 2.1

Consider the first order representation

$$\dot{\omega}(t) = -\beta\omega(t) + \beta \sum_{k=-\infty}^{\infty} e(kT_\omega)p(t-kT_\omega)$$

where the sequence $e(kT_\omega)$ is a zero-mean independent WSS process with covariance

$$\xi\{e^2(kT_\omega)\} = \frac{1 - e^{-2\beta T_\omega}}{(1-e^{-\beta T_\omega})^2}$$

for all integers k. At $t=kT_\omega$, it follows from (2.29a-b) that $\Omega(T_\omega)=1$. Within the period T_ω, the covariance $\Omega(\delta)=r_\omega(\delta,\delta)$ can be evaluated by using the equation (2.32a-b). More generally for $kT_\omega \leq t \leq (k+1)T_\omega$ and $kT_\omega \leq s \leq (k+1)T_\omega$

$$r_\omega(t,s) = e^{-\beta(t+s)} + \frac{(1-e^{-\beta t})(1-e^{-\beta s})(1-e^{-2\beta T_\omega})}{(1-e^{-\beta T_\omega})^2}$$

The covariance $\Omega(\delta)$ which is given by $r_\omega(t,s)$ for $t=s=\delta$ can be shown to satisfy

$$\Omega(\delta) \leq \Omega(T_\omega) = 1$$

for all $0 \leq \delta \leq T_\omega$ and for all non-negative β. The minimum of $\Omega(\delta)$ occurs at δ_{min} where

$$e^{-\beta\delta_{min}} = \frac{1 + e^{-\beta T_\omega}}{2}$$

$$\Omega(\delta_{min}) = \frac{1 + e^{-\beta T_\omega}}{2}$$

Note that $\delta_{min} \to 0$ as $\beta \to \infty$. Plots of covariance $\Omega(\delta)$ of $\omega(t)$ for all $kT_\omega \leq \delta \leq (k+1)T_\omega$ for some values of β are illustrated in Fig.2.3.

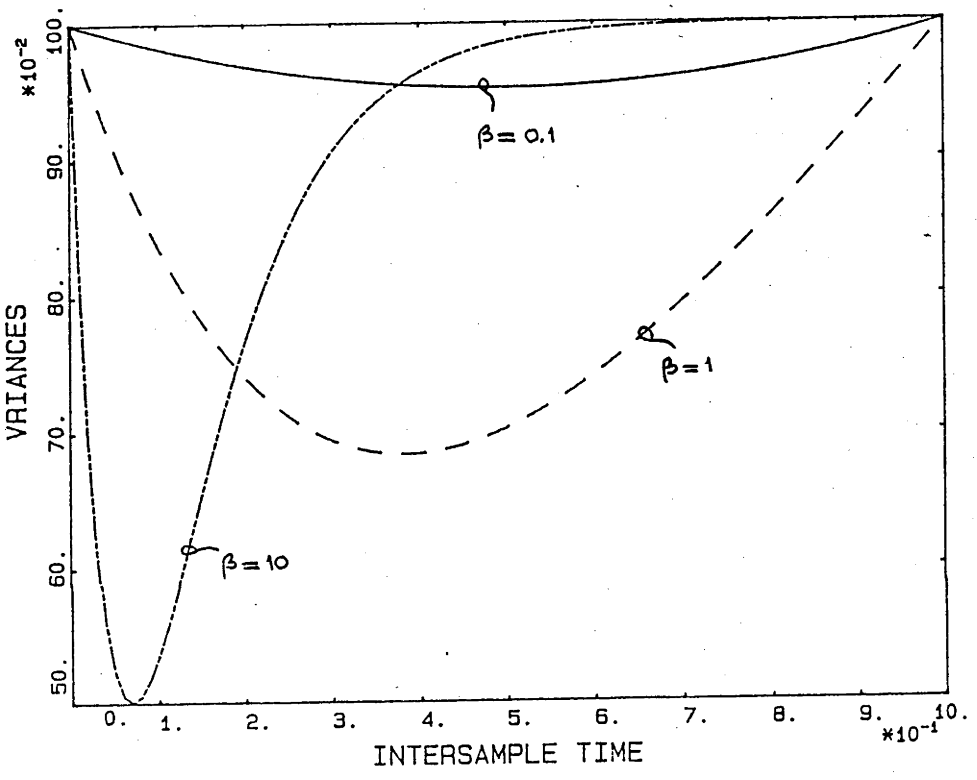
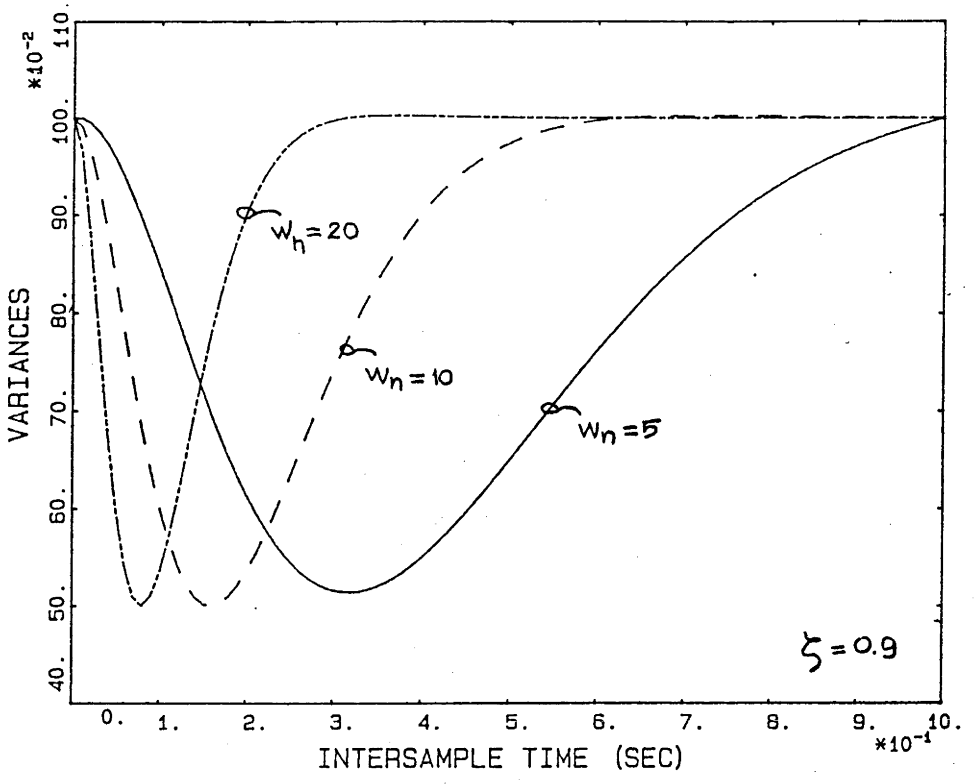


Fig.2.3 The covariance $\Omega(\delta)$ for 1st order systems

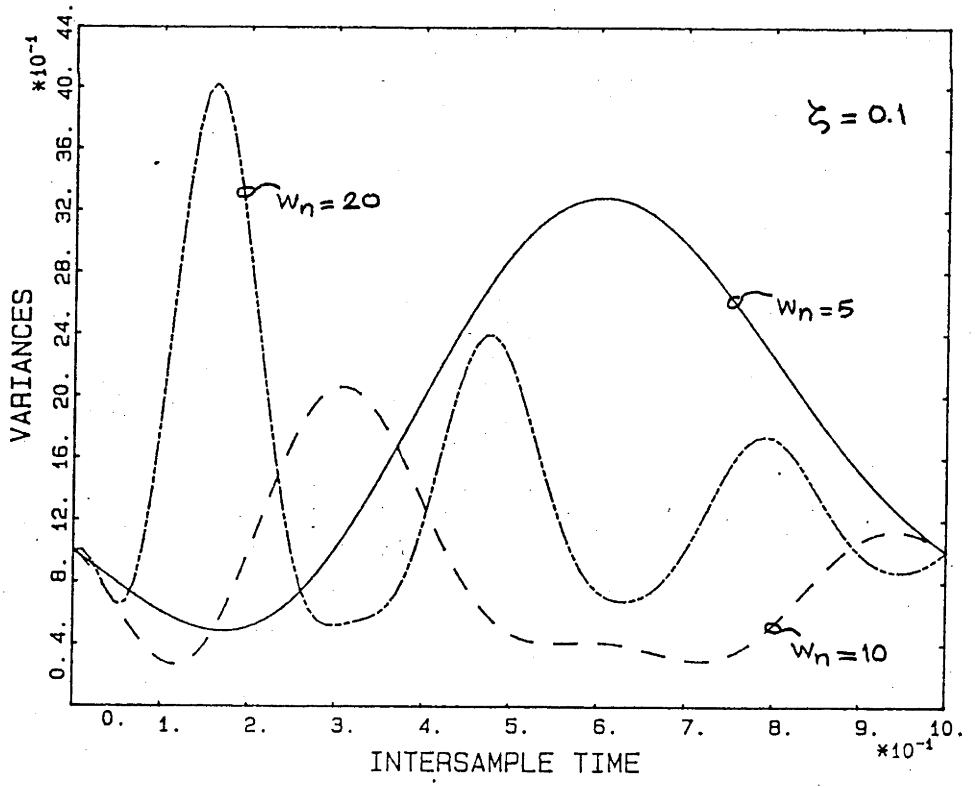
Example 2.2 Consider the second order representation

$$\ddot{\omega}(t) + 2\zeta w_n \dot{\omega}(t) + w_n^2 \omega(t) = \sum_{k=-\infty}^{\infty} e(kT_\omega) p(t - kT_\omega)$$

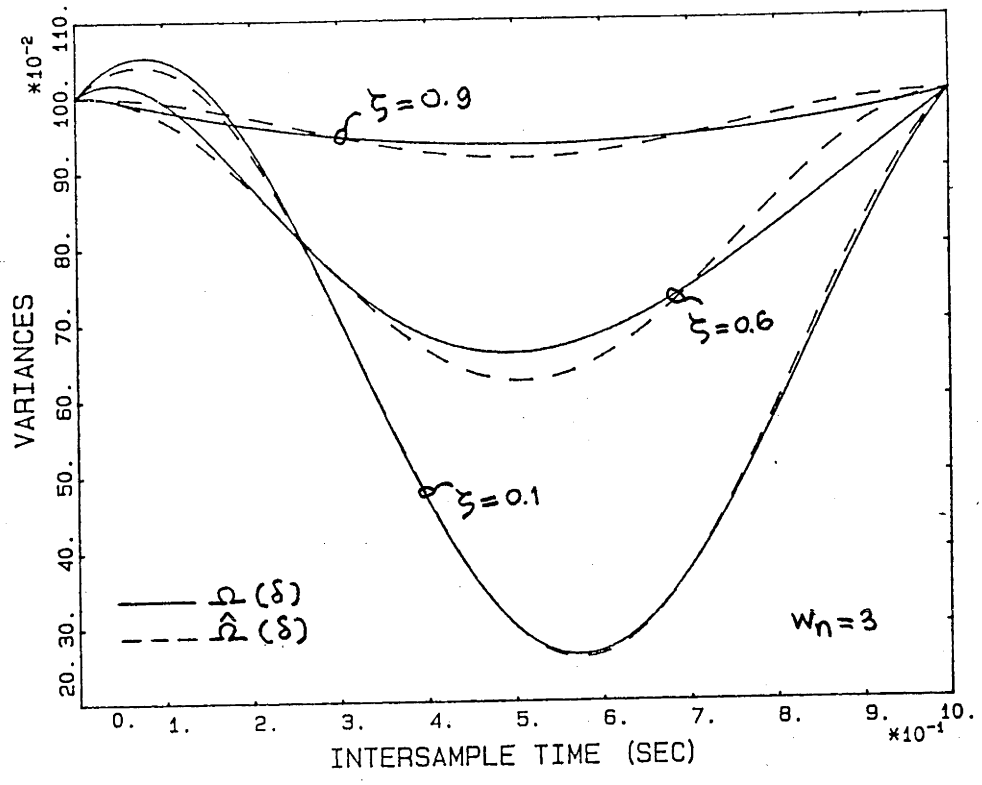
where the factors ζ and w_n are strictly non-negative and it is assumed that for each selected pair $\{\zeta, w_n\}$ the covariance $E = E(\zeta, w_n)$ of the sequence $\{e(kT_\omega)\}$ for simplicity is normalized such that the covariance $\Omega(T_\omega)=1$. The covariance $\Omega(\delta)$ can be evaluated using the covariance equations (2.29a-b) and (2.31a-b). Plots of covariance $\Omega(\delta)$ of $\{\omega(t)\}$ for all $kT_\omega \leq \delta \leq (k+1)T_\omega$ for some combination of ζ and w_n


VARIANCES
*10⁻²
110.
100.
90.
80.
70.
60.
50.
40.
ωn=20
ωn=10
ωn=5
ζ=0.9
0. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10.
*10⁻¹
INTERSAMPLE TIME (SEC)

(a)


VARIANCES
*10⁻¹
44.
40.
36.
32.
28.
24.
20.
16.
12.
8.
4.
0.
ζ=0.1
ωn=20
ωn=5
ωn=10
0. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10.
*10⁻¹
INTERSAMPLE TIME

(b)



(c)

Fig.2.4 Covariance $\Omega(\delta)$ for 2nd order systems

are illustrated in Fig.2.4. Note that for sufficiently high values of ζ and w_n, the plots of covariance $\Omega(\delta)$ in Fig.2.4a are similar to the curves of $\Omega(\delta)$ in Fig.2.3. This fact is not surprising since it is well known in linear system theory that a second order system with a high damping ratio can be approximated by a first order system with a certain time constant. For a low damping ζ and a high natural frequency w_n, the characteristics of the covariance $\Omega(\delta)$ for this second order representation can be quite complex as evidenced in Fig.2.3b whereas for a low natural frequency w_n in Fig.2.3c the charateristics of the covariance $\Omega(\delta)$ are more sinusoidal.

We now examine the harmonic series representation of cyclostationary processes.

Definition 2.2 [Ogura (1971)] A cyclostationary process $\tilde{\omega}(t)$ with period T_ω is said to have a *harmonic series representation* if there exist jointly WSS continuous-time

representors $\{a_k(t)\}$ which are bandlimited in the interval $[-1/2T_\omega, 1/2T_\omega]$ such that

$$\xi\{\left|\tilde{\omega}(t) - \sum_{k=-\infty}^{\infty} a_k(t)e^{j2\pi kt/T_\omega}\right|^2\} = 0 \qquad (2.33a)$$

for all $-\infty<t<\infty$ where $a_k(t)$ is defined by

$$a_k(t) = \int_{-\infty}^{\infty} \frac{\sin(\pi(t-\tau)/T_\omega)}{(\pi(t-\tau)/T_\omega)}\, \omega(\tau)\, e^{-j2\pi k\tau/T_\omega}\, d\tau \qquad (2.33b)$$

□□□

The autocorrelation $\tilde{r}_\omega(t,s)$ of $\tilde{\omega}(t)$ can be derived from representation (2.33a-b); that is

$$\tilde{r}_\omega(t,s) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \tilde{r}_{nm}(t-s)e^{j2\pi(nt-ms)/T_\omega} \qquad (2.34a)$$

$$\tilde{r}_{nm}(t-s) \triangleq \xi\{a_n(t)a_m^*(s)\} \qquad (2.34b)$$

where $a_m^*(s)$ denotes the complex conjugate of $a_m(s)$. Note that in (2.33a), since the components of $\tilde{\omega}(t)$ (ie. $a_k(t)\exp(-j2\pi k\tau/T_\omega)$ for all $-\infty<k<\infty$) are individually (ie. for each k) WSS but not jointly WSS (due to the exponential factors), it follows [Gardner and Franks (1975)] that the (zero-mean) process $\{\tilde{\omega}(t)\}$ is WSS if and only if the representors $\{a_k(t)\}$ defined in (2.33b) are uncorrelated.

In the context of this thesis, we consider the (real) harmonic series representation of a zero-mean cyclostationary process $\omega(t)$ as defined by

$$\omega(t) = \sum_{k=-\infty}^{\infty} [a_{1k}(t)\cos(2\pi kt/T_\omega) + a_{2k}(t)\sin(2\pi kt/T_\omega)] \qquad (2.35a)$$

where the representors $\{a_{1k}(t)\}$ and $\{a_{2k}(t)\}$ are jointly WSS. The autocorrelation $r_\omega(t,s)$ of $\omega(t)$ can be derived from representation (2.34a); that is

$$\begin{aligned} r_\omega(t,s) = 0.5 \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} [&r_{1nm}(t-s)(\cos 2\pi(nt+ms)/T_\omega + \cos 2\pi(nt-ms)/T_\omega) + \\ &r_{2nm}(t-s)(\cos 2\pi(nt-ms)/T_\omega - \cos 2\pi(nt+ms)/T_\omega) + \\ &r_{12nm}(t-s)(\sin 2\pi(nt+ms)/T_\omega + \sin 2\pi(nt-ms)/T_\omega)] \end{aligned} \qquad (2.35b)$$

where

$$r_{1nm}(t-s) \triangleq \xi\{a_{1k}(t)a_{1k}(s)\}$$

$$r_{2nm}(t-s) \triangleq \xi\{a_{2k}(t)a_{2k}(s)\}$$

$$r_{12nm}(t-s) \triangleq \xi\{a_{1k}(t)a_{2k}(s)\}$$

From (2.34b), it can be deduce that the zero-mean process $\{\omega(t)\}$ is WSS if and only if

(i). the representors $\{a_{1k}(t)\}$ and $\{a_{2k}(t)\}$ are uncorrelated; that is

$$r_{1nm}(t-s) = 0$$

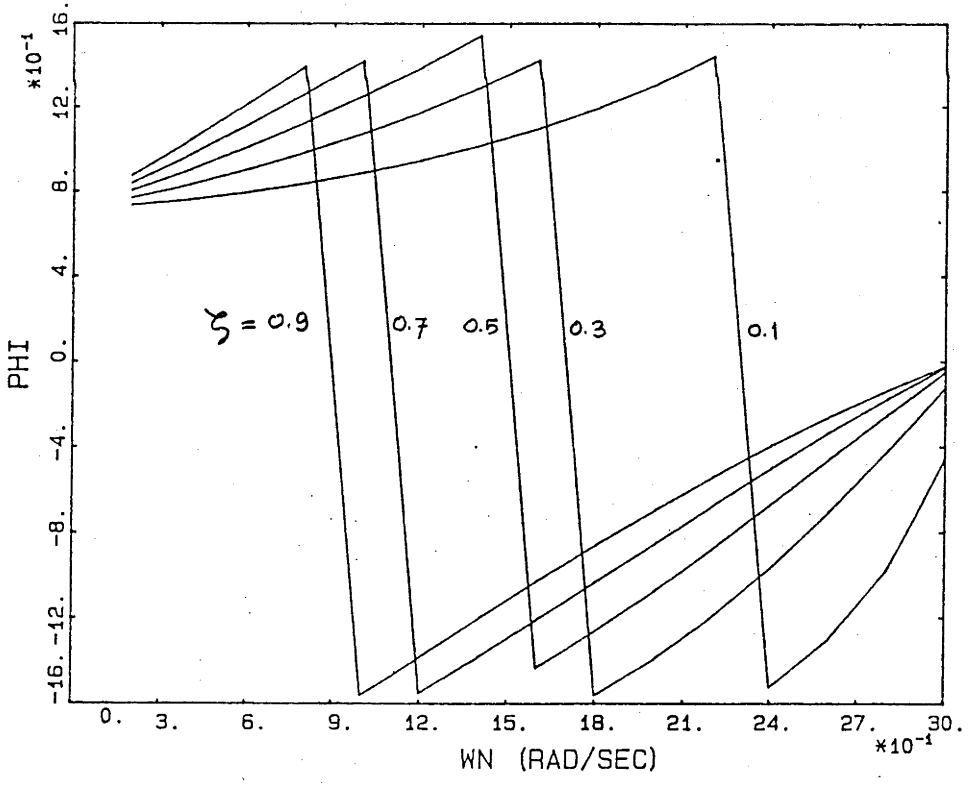
$$r_{2nm}(t-s) = 0$$

for all $t \neq s$ and $n \neq m$, and

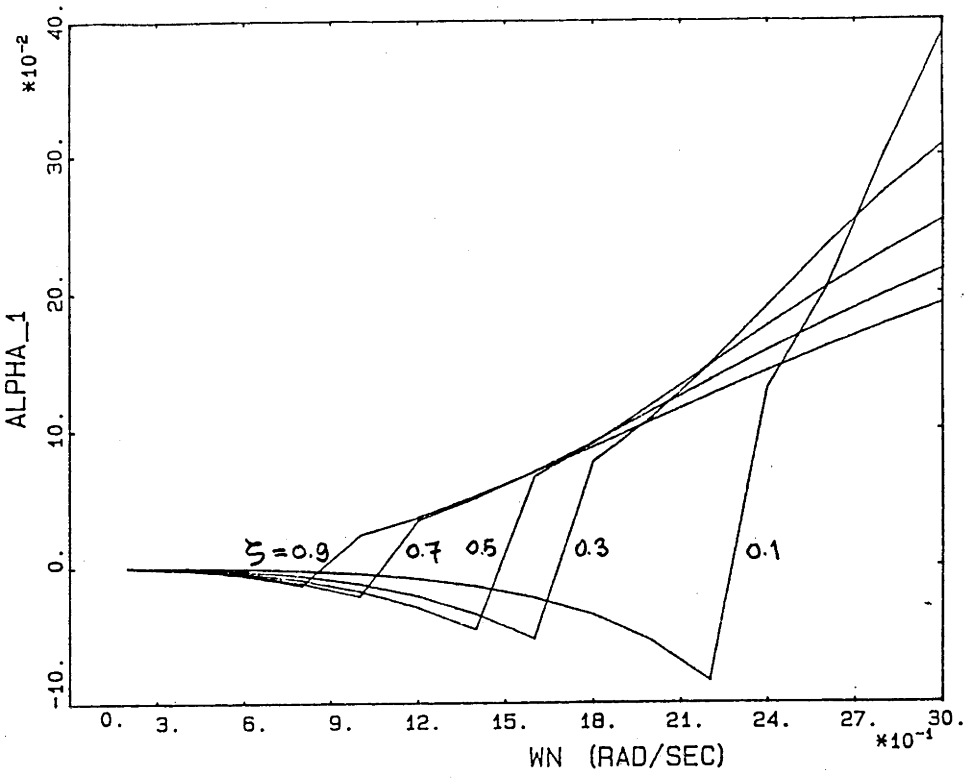
$$r_{12nm}(t-s) = 0$$

for all t,s,n and m, and

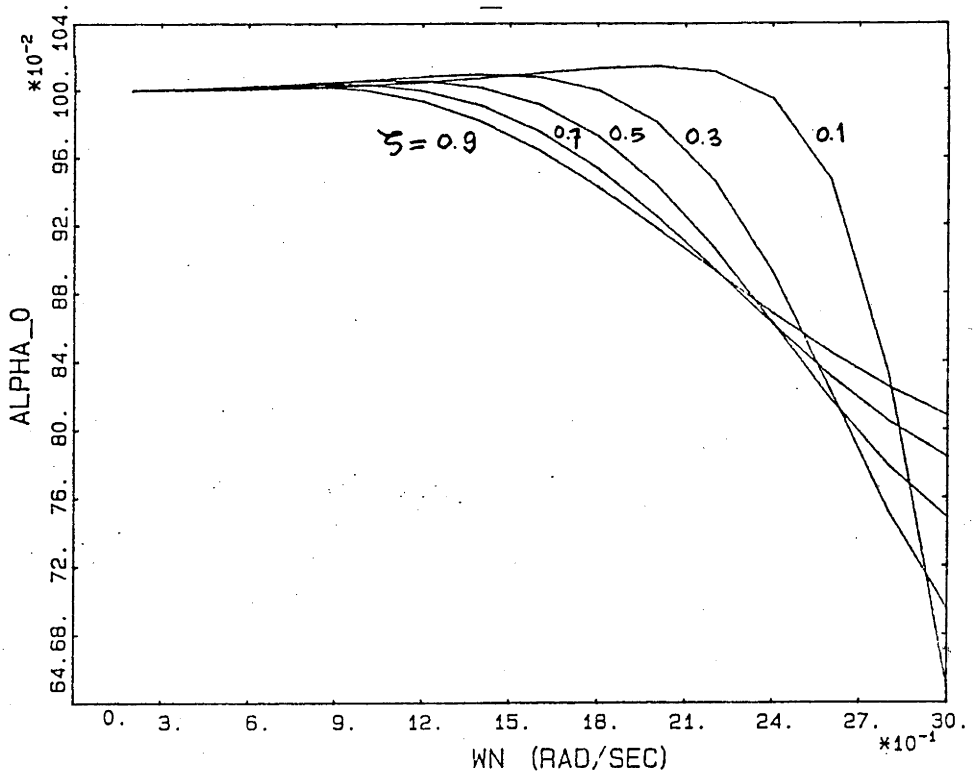
(ii). $r_{1nn}(0) = r_{2nn}(0)$



a. φ_1



b. α_1



c. α_0

Fig.2.5 Coefficients of covariance approximation $\hat{\Omega}(\delta)$

The covariance function $\Omega(\delta)=r_\omega(\delta,\delta)$ of the process $\{\omega(t)\}$ at $t=\delta$ can be derived from the autocorrelation function $r_\omega(t,s)$ in (2.35b); that is

$$\Omega(\delta) = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k \cos(2\pi k\delta/T_\omega + \varphi_k) \qquad (2.36a)$$

A first order approximation to $\Omega(\delta)$ defined in (2.32a-b) is given by

$$\hat{\Omega}(\delta) = \alpha_0 + \alpha_1 \cos(2\pi\delta/T_\omega + \varphi_1) \qquad (2.36b)$$

where α_0, α_1 and φ_1 are calculated by evaluating $\Omega(\delta)$ in (2.32a-b) by using the least square method. For the second order representation (2.27a-b), the functions α_0, α_1 and φ_1 of damping ratio ζ and natural frequency w_n are illustrated in Fig.2.5. Numerical results have revealed (see Fig.2.4c) that $\hat{\Omega}(\delta)$ is a good approximation of $\Omega(\delta)$ for $w_n < \pi T_\omega^{-1}$ and $\zeta<1$.

2.4 RELEVANT CYCLOSTATIONARY PROCESSES IN STOCHASTIC CONTROL

In the previous section, we presented series representations of cyclostationary processes namely the translational series and the harmonic series representations. In this section, we discuss the relevant cyclostationary processes in stochastic control. First, consider the continuous-time system described by (2.3a-b) where the control signal u(t) given by (2.5a-b) is an output of a zero-order sample-and-hold. We assume the control sequence $\{u(kT_c)\}$ is updated using the linear state feedback law

$$u(kT_c) = - Gx(kT_c) \qquad (2.37a)$$

The design of gain G in (2.37a) will not be discussed in this section, it will appear in chapters 4, 5 and 6. Substitution of $u(kT_c)$ in (2.7a) using (2.37a) yields

$$x((k+1)T_c) = (\Phi-\Gamma G)x(kT_c) + \omega(kT_c) \qquad (2.37b)$$

where the disturbance $\{\omega(kT_c)\}$ which is the discrete version of the continuous process $\omega(t)$ is given by (2.9a). Assuming $\omega(t)$ is WSS, it follows that the sequence $\{\omega(kT_c)\}$ is also WSS. Consequently, from (2.37b) it can be seen that the state vector $\{x(kT_c)\}$ and the output $\{y(kT_c)\}$ are also WSS. It follows from (2.37a) that

the control sequence $\{u(kT_c)\}$ is WSS. Therefore, the control signal u(t) in (2.5a-b) is a pulse amplitude modulated (PAM) signal with random amplitude. This particular signal is well known in communication applications to have periodic statistical characteristics [Franks (1969)]. The cyclostationarity of the signal u(t) which is depicted by the periodicity of the mean and the variance of u(t) can be evaluated directly from the representation (2.5a-b). The result is stated in the following lemma.

Lemma 2.3 [Franks (1969)] Consider the (scalar) control signal u(t) defined by (2.5a) with the unit pulse p(t) given in (2.5b). Assume the control sequence $\{u(kT_c)\}$ is zero mean WSS with correlation sequence $\{\rho_u(kT_c)\}$. Then the control signal u(t) is WSCS of period T_c with correlation function

$$r_u(t+\delta,t) = \sum_{k=-\infty}^{\infty} \rho_u(kT_c)q(t,\delta+kT_c) \tag{2.38a}$$

for all $0 \leq \delta < T_c$ where $q(t,\delta)$ is a periodic indicator function given by

$$q(t,\delta) = \begin{cases} 1 & \text{if } kT_c \leq t \leq (k+1)T_c - \delta \\ 0 & \text{otherwise} \end{cases} \tag{2.38b}$$

□□□

At the controlling instant $t=kT_c$, the closed loop variables $x(kT_c)$, $y(kT_c)$ and $u(kT_c)$ are all WSS. But the continuous signal u(t) as it is shown in lemma 2.3 exhibits periodic statistical properties. Consequently, the continuous state vector x(t) and output y(t) will also exhibit periodic statistical properties [de Souza and Goodwin (1984)]. To see this, consider the covariance equation (2.15a-b). At time instants $t=kT_c+n\Delta$ for a fixed n, the covariance $X(n\Delta)$ of the state vector $x(kT_c+n\Delta)$ in (2.17a) are equal for all k which implies that the process $\{x(kT_c+n\Delta):-\infty<k<\infty\}$ is stationary. However, the process is a function of n, thus the process $\{x(kT_c+n\Delta):n=0,1,2,\ldots,N\}$ defined in (2.17b) is cyclostationary. Using the same reasoning, it can be shown that the process $\{y(kT_c+n\Delta):n=0,1,2,\ldots,N\}$ is also cyclostationary. The cyclostationary characteristics of digitally controlled

continuous-time systems are presented in the following illustrative examples.

Example 2.3 Consider a scalar model (2.3a) with A=−a and B=C=1. Assume for simplicity no measurement noise (ie. Λ=0) and Ω_c=1 in (2.4a-b). The discrete version of the model (2.3a) for a sampling period T_c=1sec is governed by (2.7)−(2.9); that is

$$x(k+1) = e^{-a}x(k) + a^{-1}(1-e^{-a})u(k) + \omega(k) \tag{2.39}$$

where

$$\Omega \triangleq \xi\{\omega^2(k)\} = \frac{1-e^{-2a}}{2a} \tag{2.40}$$

In order to show the intersample characteristics of the controlled system, the control law defined in (2.6) is selected such that the variance of output y(t) at the control instants $t=kT_c$ is minimized (this is the so called minimum variance regulator (MVR) [Åström (1970)], the MVR will be discussed in chapter 4). The MVR control law for the plant (2.39) is defined by

$$u(k) = -\frac{ae^{-a}}{1-e^{-a}}x(k) \tag{2.41}$$

After substituting u(k) in (2.39) using (2.41), the covariance X(NΔ) of the state x(k+1) can be written as

$$X(N\Delta) = \Omega \tag{2.42}$$

where Ω is the covariance of discrete noise $\{\omega(k)\}$. Note that in (2.42) $N\Delta=T_c$ (where T_c=1sec). The intersample variance can be evaluated by means of lemma 2.2. From the covariance equation (2.15a), the intersample variance can be written as

$$X(n\Delta) = \left(e^{-an\Delta} - \frac{(1-e^{-an\Delta})e^{-a}}{(1-e^{-a})}\right)^2\Omega + \frac{1-e^{-2an\Delta}}{2a} \tag{2.43}$$

Fig.2.6 shows the characteristics of X(nΔ) for some values of a. From (2.43), it can be shown that for all stable a (ie. a>0) the maximum variance occurs at the controlling instants $t=kT_c$. This is also evidenced in Fig.2.6.

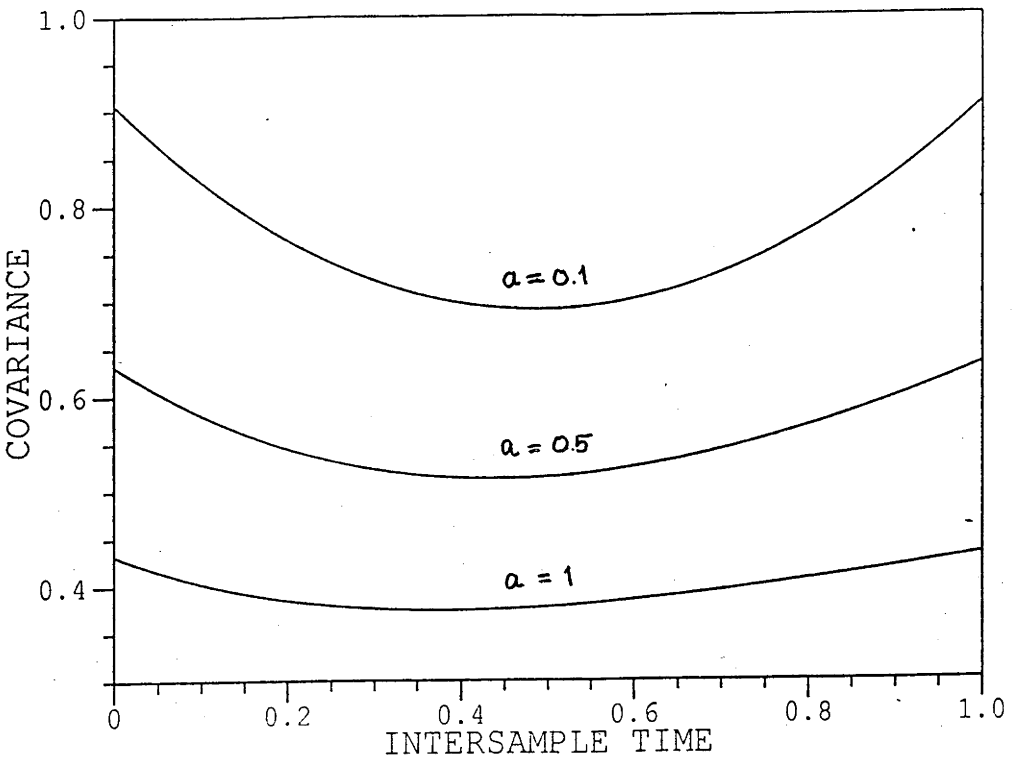


Fig.2.6 Intersample covariance $X(n\Delta)$ of 1st order system

Example 2.4 [de Souza and Goodwin (1984)] Consider a single-input single-output (SISO) second order plant (2.3a-b) where the system matrices A, B and C are given as follows

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -0.1 \end{bmatrix} \quad ; \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad ; \quad C' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\Omega_c = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad ; \quad \Lambda = 0$$

As in example 2.3, the selection of the control sequence $\{u(kT_c)\}$ defined in (2.6) is based on the MVR criterion. The MVR gains denoted by G in (2.6) for different values of sampling period T_c are given as follows.

i. T_c = 1 sec. ; G = [2.0672 1.9672]

ii. T_c = 0.5 sec. ; G = [8.1339 3.9669]

iii. T_c = 0.1 sec. ; G = [200.6672 19.9667]

T_c(sec)	ratio	worst variance
1	12.81	3.9653
0.5	25.02	1.0037
0.1	122.51	0.0405

Table 2.1 The effects of sampling on intersample variance of a 2nd order system

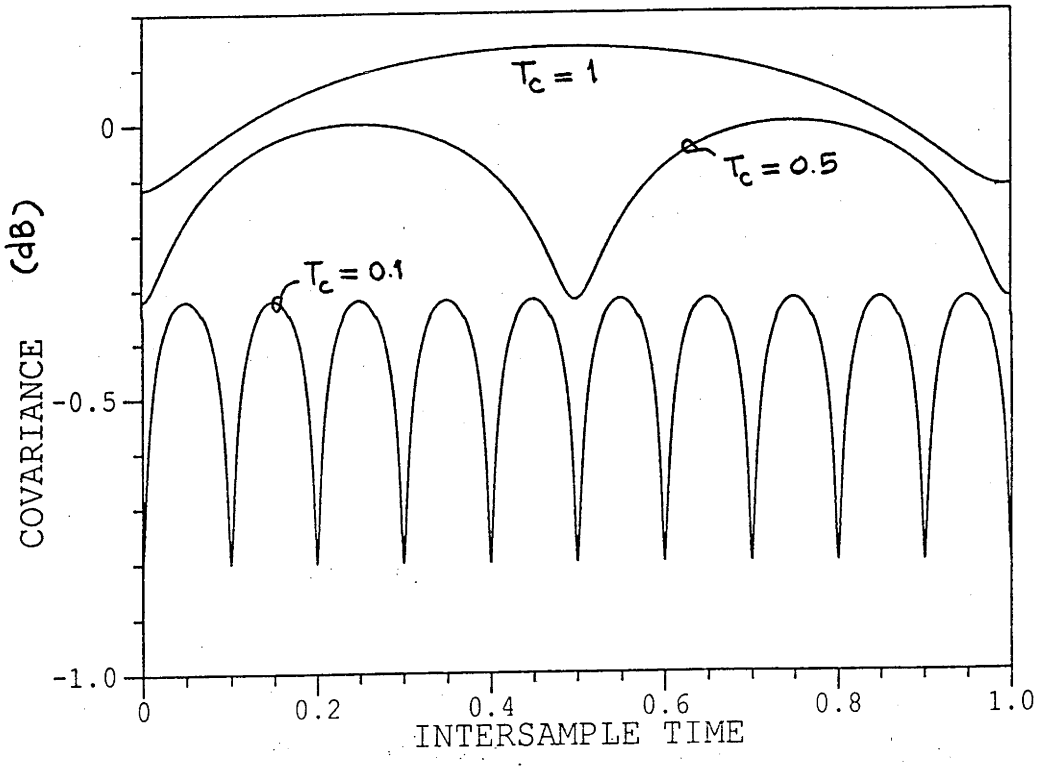


Fig.2.7 Intersample output variance of a 2nd order system

The resulting intersample output variance is depicted in Fig.2.7. From Fig.2.7 and Table 2.1, it can be seen that the ratio between the worst intersample variance and the variance at the control instants $t=kT_c$ increases with the sampling rate T_c^{-1}.

Example 2.3 Consider again the SISO second order model (2.3a-b) where the matrices A and B are given by

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad ; \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

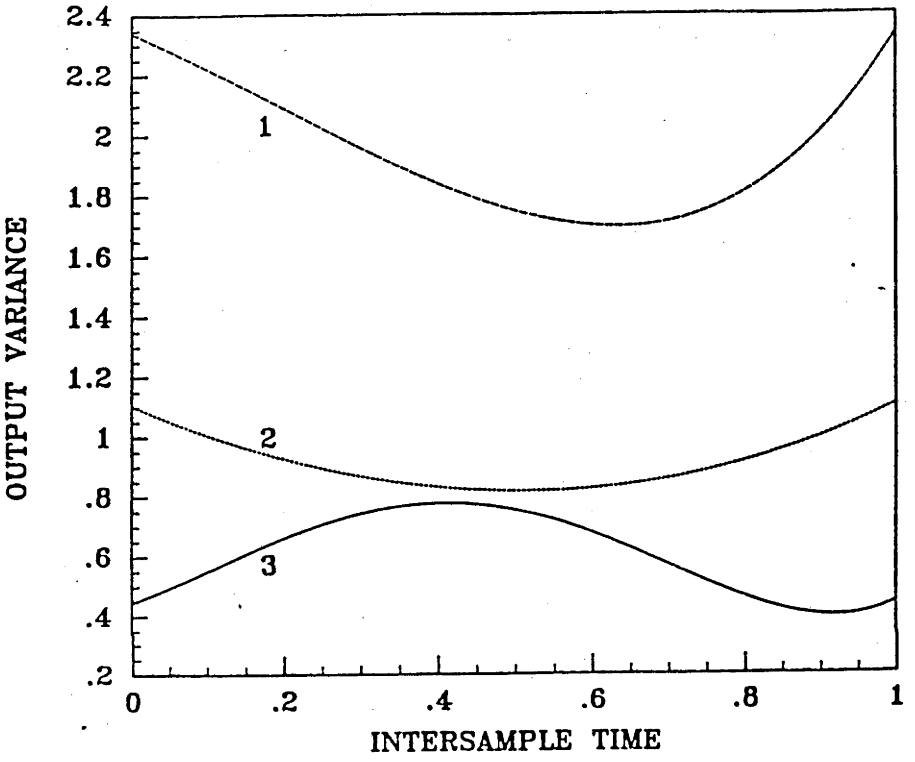


Fig.2.8 Intersample output variance of 2nd order systems

Assume the sampling period is fixed at T_c=1sec. The output of the model is considered for three different choices of output vector C and for each given values of C we assume the state feedback gain G is selected to satisfy the MVR criterion mentioned in the example 2.3. We list the output vectors C and the gains G as follows.

i. C = [0.1 1] ; G = [0.6667 1.3333]

ii. C = [1 1] ; G = [1.6667 1.8833]

iii. C = [1 0.1] ; G = [0.0952 1.0476]

The intersample characteristics of the output corresponding to each pair {C,G} were computed by means of lemma 2.2 and the results are illustrated in Fig.2.8. It can be seen from Fig.2.8 that the intersample statistical response of the controlled system for a certain pair {C,G} is different to that of the controlled system for another pair {C,G}.

So far, we have only considered the case where the process disturbance and the measurement noise are WSS. In the following lemma, we present the characteristics of digitally controlled continuous-time systems in the presence of both WSS and WSCS disturbances.

Lemma 2.4 Consider the continuous-time SISO system (2.3a-b) with the linear state feedback control law u(t) defined by (2.5a-b) where the processes $\{\omega(t)\}$ and $\{\dot{\eta}(t)\}$ are zero mean independent WSS and are uncorrelated with the state x(t), the input u(t) and the output y(t). Then, if the control sequence $\{u(kT_c)\}$ is zero mean WSS, the continuous processes {x(t)} and {y(t)} are zero mean WSCS of period T_c. More generally, if the process $\{\omega(t)\}$ is WSCS of period T_ω with $mT_c=nT_\omega$ where n and m are integers and the ratio n/m is irreducible, then the continuous processes {x(t)} and {y(t)} are WSCS of period mT_c $(=nT_\omega)$.

Proof: From lemma 2.3, the control signal u(t) defined by (2.5a) is WSCS with period T_c. Suppose the impulse response function of the system (2.3a) is given by h(t) where

$$h(t) = \begin{cases} Ce^{At}B & \text{for } t\geq 0 \\ 0 & \text{for } t<0 \end{cases}$$

The influence of the input u(t) on the output y(t) can be expressed as

$$y(t) = \int_{-\infty}^{\infty} h(t-\sigma)u(\sigma)d\sigma$$

The mean function of the output $y(t+T_c)$ is given by

$$\xi\{y(t+T_c)\} = \int_{\infty} h(t+T_c-\sigma)\xi\{u(\sigma)\}d\sigma$$

$$= \int_{-\infty}^{\infty} h(t-\nu)\xi\{u(\nu+T_c)\}d\sigma$$

but $\xi\{u(\nu+T_c)\}=\xi\{u(\nu)\}=0$ and hence

$$\xi\{y(t+T_c)\} = \xi\{y(t)\} = 0$$

The autocorrelation function of the output $y(t+T_c)$ is given by

$$r_y(t+T_c,s+T_c) = \xi\{y(t+T_c)y'(s+T_c)\}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t+T_c-\sigma)r_u(\sigma,\tau)h'(s+T_c-\tau)d\sigma d\tau$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\gamma)r_u(\gamma+T_c,\nu+T_c)h'(s-\nu)d\gamma d\nu$$

The cyclostationarity of the control signal u(t) implies $r_u(\gamma+T_c,\nu+T_c)=r_u(\gamma,\nu)$ and hence

$$r_\omega(t+T_c,s+T_c) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\gamma)r_u(\gamma,\nu)h'(s-\nu)d\gamma d\nu$$

$$= r_\omega(t,s)$$

The periodicity of the mean function $\xi\{y(t)\}$ and of the autocorrelation function $r_\omega(t,s)$ establish the cyclostationarity of the output y(t) (under the influence of both u(t) and $\omega(t)$ (which is WSS)). Using a similar analysis, the cyclostationarity of the state x(t) defined by (2.3a) can also be established. To prove the more general result, let the output y(t) be decomposed as

$$y(t) = y_1(t) + y_2(t)$$

where $y_1(t)$ is the response due to the input u(t) and $y_2(t)$ is the output due to disturbance $\omega(t)$. Note that $y_1(t)$ and $y_2(t)$ are uncorrelated since u(t) and $\omega(t)$ are uncorrelated. The output y(t) is related to u(t) and $\omega(t)$ by

$$y(t) = \int_{-\infty}^{\infty} h(t-\sigma)u(\sigma)d\sigma + \int_{-\infty}^{\infty} ce^{A(t-\sigma)}\omega(\sigma)d\sigma$$

where $h(t-\sigma)$ is the impulse response function of the system (2.3a) and C and A are

defined in (2.3a). The inputs u(t) and $\omega(t)$ are both zero mean. Therefore, the mean function $\xi\{y(t)\}$ is given by

$$\xi\{y(t+mT_c)\} = \xi\{y(t)\} = 0$$

where $mT_c=nT_\omega$. The autocorrelation function $r_\omega(t+mT_c,s+mT_c)$ of the output y(t) is given by

$$r_\omega(t+mT_c,s+mT_c) = \xi\{y(t+mT_c)y'(s+mT_c)\}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t+mT_c-\sigma)r_u(\sigma,\tau)h'(s+mT_c-\tau)d\sigma d\tau +$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Ce^{A(t+mT_c-\sigma)}r_\omega(\sigma,\tau)e^{A'(s+mT_c-\tau)}C' d\sigma d\tau$$

where h(t) is the impulse response function (from the input u(t)) of the system (2.3a) and C and A are defined in (2.3a). Let $\sigma=\gamma+mT_c$ and $\nu=\tau+mT_c$, we have

$$r_\omega(t+mT_c,s+mT_c) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\gamma)r_u(\gamma+mT_c,\nu+mT_c)h'(s-\nu)d\gamma d\nu +$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Ce^{A(t-\gamma)}r_\omega(\gamma+mT_c),\nu+mT_c)e^{A'(s-\nu)}C' d\gamma d\nu$$

The cyclostationarity of u(t) (of period T_c) and $\omega(t)$ (of period T_ω where $mT_c=nT_\omega$) imply

$$r_u(\gamma+mT_c,\nu+mT_c) = r_u(\gamma,\nu)$$

$$r_\omega(\gamma+mT_c,\nu+mT_c) = r_\omega(\gamma+\nu)$$

and hence

$$r_y(t+mT_c,s+mT_c) = r_y(t,s)$$

The periodicity of the mean function $\xi\{y(t)\}$ and of the autocorrelation function $r_y(t,s)$ establish the cyclostationarity of the output y(t). Using a similar analysis the state x(t) can also be shown to be WSCS with period $nT_\omega(=mT_u)$. This completes the proof.

Remark: Suppose the control law $u(kT_c)$ is assumed to be a state feedback control law where the (true) state is available for control. Then, if only the measurement

noise $\eta(t)$ in (2.3b) is WSCS of period T_η the state x(t) will not be WSCS.

□□□

Another cyclostationary process that may occur in control application, is a cyclostationary measurement noise. In the electronic communication this type of non-stationary process is not at all uncommon, see for example [Franks (1969)]. It usually appears as a consequence of periodic operations (such as sampling an scanning) which are intentionally introduced to transform the (received) signal into a prescribed format. In this section, we consider the (scalar) measurement signal of the form

$$y(t) = Cx(t) + s(t) \tag{2.44a}$$

where the disturbance s(t) is governed by

$$s(t) = \sum_{q=1}^{N} [\alpha_q(t) \cos w_q t + \beta_q(t) \sin w_q t] \tag{2.44b}$$

In the deterministic case, that is when $\alpha_q(t)$ and $\beta_q(t)$ are constant but unknown for all q, this type of sinusoidal disturbance may be produced by some rotational parts in the physical system. For example, in [Goodwin et. al (1986)] it is shown that s(t) in (2.44b) is the disturbance due to the rotor motion of a helicopter. A more general type of disturbance is formed if $\alpha_q(t)$ and $\beta_q(t)$ in (2.44b) are assumed to be WSS processes with the following characteristics

$$p(t) \triangleq [\alpha_1(t)\ \beta_1(t)\ \alpha_2(t)\ \beta_2(t)\ \ldots\ \alpha_N(t)\ \beta_N(t)]' \tag{2.45a}$$

$$\xi\{p'(t)\} = [m(\alpha_1)\ m(\beta_1)\ m(\alpha_2)\ m(\beta_2)\ \ldots\ m(\alpha_N)\ m(\beta_N)] \tag{2.45b}$$

$$\xi\{p(t)p'(s)\} = 0 \quad \text{for } t \neq s \tag{2.45c}$$

$$m(\alpha_i) \triangleq \xi\{\alpha_i(t)\} \tag{2.45d}$$

$$V(\alpha_i) \triangleq \xi\{\alpha_i^2(t)\} \tag{2.45e}$$

$$V(\beta_i) \triangleq \xi\{\beta_i^2(t)\} \tag{2.45f}$$

$$V(\alpha\beta_i) \triangleq \xi\{\alpha_i(t)\beta_i(t)\} \tag{2.45g}$$

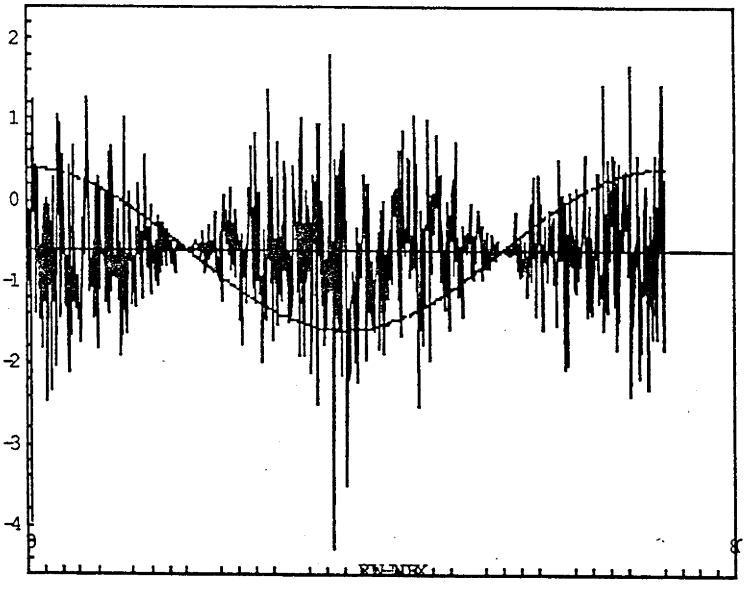


Fig.2.9 A simple example of a cyclostationary disturbance

$$\xi\{p(t)p^{*}(t)\} =$$

$$\begin{bmatrix}
V(\alpha_1) & V(\alpha\beta_1) & 0 & 0 & \cdots & 0 & 0 \\
V(\alpha\beta_1) & V(\beta_1) & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & V(\alpha_2) & V(\alpha\beta_2) & \cdots & 0 & 0 \\
0 & 0 & V(\alpha\beta_2) & V(\beta_2) & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & V(\alpha_N) & V(\alpha\beta_N) \\
0 & 0 & 0 & 0 & \cdots & V(\alpha\beta_N) & V(\beta_N)
\end{bmatrix} \qquad (2.45h)$$

for all $i\epsilon[1,N]$.

When N=1 and $\beta_q(t)=0$, the disturbance s(t) in (2.44b) can be written as

$$s(t) = \alpha_1(t)\cos w_1 t \qquad (2.46)$$

Fig.2.9 shows a typical realization of the disturbance s(t) in (2.46). The first and

second moments of the signal s(t) in (2.44b) are presented in the following lemma.

Lemma 2.5 Consider the cyclostationary disturbance s(t) defined by (2.44b) where the processes $\{\alpha_q(t)\}$ and $\{\beta_q(t)\}$ are assumed to be WSS, and are uncorrelated with the characteristics given by (2.45a-h). Then, the first and the second moments of the disturbance s(t) are governed by

$$m_s(t) \triangleq \xi\{s(t)\}$$

$$= \sum_{q=1}^{N} [m(\alpha_q)\cos w_q t + m(\beta_q)\sin w_q t] \tag{2.47a}$$

$$V_s(t) \triangleq \xi\{s^2(t)\}$$

$$= 0.5\{\sum_{q=1}^{N} [r_0(q) + r_1(q)\cos 2w_q t + 2r_2(q)\sin 2w_q t]\} \tag{2.47b}$$

where

$$r_0(i) = V(\alpha_i) + V(\beta_i) \tag{2.47c}$$

$$r_1(i) = V(\alpha_i) - V(\beta_i) \tag{2.47d}$$

$$r_2(i) = V(\alpha\beta_i) \tag{2.47e}$$

for all $i\epsilon[1,N]$ and where $m(\alpha_i)$, $V(\alpha_i)$, $V(\beta_i)$ and $V(\alpha\beta_i)$ are defined by (2.45d-g).

Proof: The first moment $m_s(t)$ of the signal s(t) in (2.47a) can be directly derived by taking the mathematical expectation of the disturbance s(t) in (2.44b). The second moment $V_s(t)$ of the process {s(t)} can be expressed as follows

$$V_s(t) = \xi\{s^2(t)\}$$

$$= \sum_{q=1}^{N} V(\alpha_q)\cos^2 w_q t + V(\beta_q)\sin^2 w_q t + 2V(\alpha\beta_q)\cos w_q t \sin w_q t$$

$$= \sum_{q=1}^{N} 0.5V(\alpha_q)(1+\cos 2w_q t) + 0.5V(\beta_q)(1-\cos 2w_q t) + V(\alpha\beta_q)\sin 2w_q t \tag{2.48}$$

A little arrangement of (2.48) and use the relations $r_0(i)$, $r_1(i)$ and $r_2(i)$ for all $1 \leq i \leq N$ defined in (2.47c-e) yield the second moment $V_s(t)$ in (2.47b).

□□□

2.5 CONCLUSIONS

The translational and harmonic series representations of the WSCS processes have been investigated in this chapter. A state space translational representation of WSCS processes has been developed and used to characterize two classes of first and second order processes. It has been shown by means of examples that the characteristics of a second order model with high damping ratio ζ are similar to the characteristics of a first order model where the time constant $\beta=0.5w_n$ where w_n is the natural frequency of the second order model. This is line with the deterministic properties of first and second order models. For a harmonic series representation, numerical results have revealed (see Fig.2.4c) that the covariance $\hat{\Omega}(\delta)$ in (2.35b) is a good approximation of the covariance $\Omega(\delta)$ in (2.36a) for a (relatively) low natural frequency (ie. $\omega_n<\pi T_\omega^{-1}$) and small damping ratio (ie. $\zeta<1$).

It has been shown that the continuous time statistical response of digitally controlled linear time-invariant systems exhibit periodic characteristics known as cyclostationary characteristics. It can be seen from the examples presented in section 3.4 that the higher the sampling (or the control rate) the larger the ratio between the (worst) intersample variance and the variance observed at the controlling instants. This phenomenon is more apparent for lightly-damped open-loop (ie. $\xi<<1$) systems. In the deterministic case, the problem of intersample dynamics has been recognized since the early development of the discrete control system [Tou (1959)]. The functional form of the intersample variance of a digitally controlled system may be quite complex (see for example the curves depicted in Fig.2.3 and Fig.2.4). A method of predicting the level of (deterministic) intersample ripple in closed-loop discrete systems which is based on the knowledge of the plant transfer function and sample-and-hold device has been outlined in [Quinn and Williamson (1985)]. But to our knowledge no similar results have been presented for the stochastic control application.

The properties of WSCS measurement noise which usually occurs as a consequence of periodic operations such as pulse amplitude modulation (PAM) in the

system has been explored in section 2.4. It has been shown that each component of the first order moment $m_s(t)$ in (2.47a) and each deterministic component of the disturbance s(t) (ie. $\cos w_q t$ and $\sin w_q t$) in (2.44b) have equal period while for the second moment $v_s(t)$ in (2.47b) the period is halved. When the functions $\alpha_q(t)$ and $\beta_q(t)$ are constant for all t but unknown, the process (2.44b) is known as a sinusoidal disturbance [Goodwin et. al (1986)]. The problem of estimating the constants α_q and β_q for $q \epsilon [1,p]$ where the integer p is known has been discussed in [Goodwin et. al (1986), Bitmead et. al (1986), Williamson (1987)].

CHAPTER 3

OPTIMAL STATE PREDICTION

3.1 INTRODUCTION

In a linear time-invariant state-space realization, the (measurable) output of a plant is usually a linear combination of the states with an injection of a random disturbance (or noise). The classical filtering problem concerns the separation of the desired output from the noise. When the frequency spectrum of the output and the noise is not overlapping, the output can be extracted from the noise by filtering. The frequency spectrum of the required filter is determined by the relative frequency of the output and the noise, it could be a low-pass, high-pass, band-pass or band-stop filter [Rabiner and Gold (1975)]. The filtering problem that we consider in this chapter concerns the extraction of the output (or the states) from the noise when the frequency spectrum of the output and the noise are overlapping. This particular problem was first studied by [Kolmogorov (1941), Wiener (1949)]; the statistical properties of the output and the noise are assumed to be constant (ie. independent of time); in other words the output and the noise are assumed to be stationary processes. It has been shown that the statistical properties of the output and the (unwanted) noise are related to their frequency domain properties. About a decade later, the so called *Kalman filter* theory which does not require the stationarity assumption was developed [Kalman (1960-3), Kalman and Bucy (1961), Kailath (1974)].

We now briefly discuss the conventional Kalman filter design problem. Consider a minimal continuous-time system described by

$$\dot{x}(t) = Ax(t) + Bu(t) + \omega(t) \tag{3.1a}$$

$$y(t) = Cx(t) + \tilde{\eta}(t) \tag{3.1b}$$

where the dimensions of the state vector x(t) and the output y(t) are given by

$$x(t)\epsilon R^{n_x} \quad ; \quad y(t)\epsilon R^{n_y} \quad ; \quad u(t)\epsilon R^{n_u}$$

The dimension of the matrices A, B and C are respectively $(n_x x n_x)$, $(n_x x n_u)$ and $(n_y x n_x)$. The processes $\{\omega(t),\tilde{\eta}(t):-\infty<t<\infty\}$ are assumed to be zero-mean independent *wide sense stationary* (WSS) with covariance

$$\xi\{\begin{bmatrix}\omega(t)\\ \tilde{\eta}(t)\end{bmatrix}[\omega'(t)\ \tilde{\eta}'(t)]\} = \begin{bmatrix}\Omega_c & \psi_c\\ \psi_c' & \Lambda_c\end{bmatrix} \tag{3.2}$$

For simplicity, we assume the cross-term $\psi_c=0$ (when $\psi_c\neq 0$, the resulting Kalman filter design requires only a little modification [Goodwin and Sin (1984)]). We also assume that the output y(t) in (3.1b) is passed through an (ideal) pre-filter which passes the term Cx(t) unchanged and only filter the wide-band noise $\tilde{\eta}(t)$ defined in (3.1b). The pre-filtered noise is denoted by $\eta(t)$.

As described in chapter 2, the discrete-time equivalent of the continuous-time model (3.1a-b) for a sampling period T_c is given by

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \tag{3.3a}$$

$$y(kT_c) = Lx(kT_c) + \eta(kT_c) \tag{3.3b}$$

where $\eta(kT_c)$ is the sampled version of the pre-filtered noise $\eta(t)$, and where

$$\Phi = e^{AT_c}$$

$$\Gamma = \int_0^{T_c} e^{A(T_c-\sigma)}B\,d\sigma$$

$$L = C$$

The sampling period T_c is selected such that the plant (3.3a-b) is also minimal. The covariance of the discrete processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$ are given by

$$\xi\{\begin{bmatrix}\omega(kT_c)\\ \eta(kT_c)\end{bmatrix}[\omega'(kT_c)\ \eta'(kT_c)]\} = \begin{bmatrix}\Omega & 0\\ 0 & \Lambda\end{bmatrix} \tag{3.4a}$$

where Ω and Λ are given by

$$\Omega = \int_0^{T_c} e^{A(T_c-\sigma)}\Omega_c e^{A'(T_c-\sigma)}\,d\sigma \tag{3.4b}$$

$$\Lambda = \frac{\Lambda_c}{T_c} \tag{3.4c}$$

Assume that the measurements (ie. data)

$$\{y(kT_c): \ldots, k-2, k-1, k\} \tag{3.5}$$

are available. Using the data (3.5), we intend to reconstruct (or to estimate) the state $x((k+n)T_c)$ for a certain integer n. There are three categories of filtering problems [Anderson and Moore (1979)] depending on the choice of n; namely

a. *smoothing*, ie. $n<0$

b. *filtering*, ie. $n=0$

c. *prediction*, ie. $n>0$

We restrict the investigation to the prediction problem. The one-step ahead predictor [Anderson and Moore (1979), Gelb (1974)] can be written as

$$\hat{x}((k+1)T_c) = \Phi\hat{x}(kT_c) + \Gamma u(kT_c) + K(y(kT_c)-\hat{y}(kT_c)) \tag{3.6a}$$

$$\hat{y}(kT_c) = L\hat{x}(kT_c) \tag{3.6b}$$

where $\hat{x}(kT_c)$ (which is a simplified notation for $\hat{x}(kT_c|(k-1)T_c)$ is the conditional mean of the state $x(kT_c)$. Define the prediction error

$$\varepsilon(kT_c) \triangleq x(kT_c) - \hat{x}(kT_c) \tag{3.7a}$$

The prediction error equation can be derived from the discrete model (3.3a-b) and the predictor equation (3.6a-b). Thus,

$$\varepsilon((k+1)T_c) = \Phi\varepsilon(kT_c) + \omega(kT_c) - K(y(kT_c)-\hat{y}(kT_c)) \tag{3.7b}$$

The state prediction problem can be defined as

$$\min_{\{K\}} \xi\{\varepsilon(kT_c)\varepsilon'(kT_c)\} \tag{3.7c}$$

Furthermore, the minimization problem (3.7c) is achieved via the minimization

$$\min_{\{K\}} tr(P) \tag{3.7d}$$

where P is the covariance of the prediction error $\varepsilon(kT_c)$ and tr(.) denotes trace of

a matrix. The optimal Kalman filter gain K is given by

$$K = \Phi P L' (LPL'+\Lambda)^{-1} \tag{3.8a}$$

where the matrix P which is in fact the covariance of the prediction error $\varepsilon(kT_c)$ in (3.7a) satisfies the algebraic Riccati equation (ARE)

$$P = \Phi P \Phi' + \Omega - \Phi P L' (L' PL+\Lambda)^{-1} LP\Phi' \tag{3.8b}$$

The ARE in (3.8b) has received much attention [Martensson (1971), Kucera (1972), Chan et. al (1984)]. We are interested in the *stabilizing solution* P (ie. when P is real, symmetric and positive semi-definite and when the eigenvalues $|\lambda_i(\Phi-KL)|<1$ for all $i\epsilon[1,n_x]$) of the ARE in (3.8b). The following lemma establishes such solution.

Lemma 3.1 [Chan et. al (1984)] Consider the minimal discrete-time model (3.3a-b) and consider also the Kalman filter (3.6a-b) and the corresponding algebraic Riccati equation (3.8b). Suppose the covariance matrix Ω (≥ 0) defined in (3.4a-b) is factorized as follows

$$\Omega = DD' \tag{3.9}$$

Then, if the pair $\{\Phi,D\}$ has no uncontrollable modes on the unit circle, the stabilizing solution of the ARE in (3.8b) exists and is unique.

□□□

Note that in (3.9), because $\Omega \geq 0$ there exists an infinity of matrices D (and all satisfy (3.9)). We assume throughout the thesis that the stabilizing solution always exists.

In this chapter, we examine the state prediction problem when the processes $\{\omega(t)\}$ and $\{\eta(t)\}$ in (3.1a-b) are *wide sense cyclostationary* (WSCS) with period T_ω and T_η. First, we examine the case when the control sampling (ie. when the control law $u(kT_c)$ is updated) and the measurement sampling (ie. when the output $y(kT_c)$ is measured) are *non-synchronous*. We then continue with the investigation of *multirate* sampling (ie. when the control and the measurement rates are different).

In section 3.2, we assume that the measurement and control instants are non-synchronous. The consequences of cyclostationary disturbances in state prediction

are then explored. We show that the optimal prediction is not generally achieved by synchronous control and output sampling. In practice, digital compensators often employ different sampling periods. Such systems are called multirate digital control systems [Glasson (1983), Brousard and Glasson (1980) and Brousard et. al (1985)]. In section 3.3, we consider different control and measurement sampling rates. We investigate the consequences of cyclostationary noises in the state reconstruction when the plant output is measured at a faster rate than the control change; that is,

$$T_c = NT_m \tag{3.10}$$

for a (strictly) positive integer N where T_c and T_m are respectively the control and measurement sampling periods. This implies that there are N measurements available for constructing a full state vector $\hat{x}(kT_c)$. This situation may arise when the on-board computer requires T_c seconds to complete the calculation for update of the control sequence $\{u(\ell T_c)\}$ while the measurement can be done at each instant $t_m=kT_m$. We show that the multirate prediction problem is similar to the non-synchronous state prediction problem in the sense that the optimal predictor is achieved by introducing a 'delay' between the measurement and the control instants.

3.2 NON-SYNCHRONOUS STATE PREDICTION

Consider the minimal SISO continuous-time system (3.1a-b) but now we assume the processes $\omega(t)$ and $\eta(t)$ and uncorrelated, and are zero mean WSCS with period T_ω and T_η with respective covariances

$$r_\omega(t,t) \triangleq \Omega_c(t)$$
$$= r_\omega(t+T_\omega, t+T_\omega) \tag{3.11a}$$
$$r_\eta(t,t) \triangleq \Lambda_c(t)$$
$$= r_\eta(t+T_\eta, t+T_\eta) \tag{3.11b}$$

At the controlling instants $t=kT_c$, the discrete-time equivalent description of the plant (3.1a-b) are given by (3.3a-b). The covariance of the discrete processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$ are given by

$$\Omega(T_c) = \int_0^{T_c}\int_0^{T_c} e^{A(T_c-\tau)} r_\omega(\sigma,\tau) e^{A'(T_c-\tau)} d\sigma d\tau \qquad (3.12a)$$

$$\Lambda(T_c) = \frac{\Lambda_c(T_c)}{T_c} \qquad (3.12b)$$

where $\Lambda_c(T_c)$ is defined in (3.11b) for $t=T_c$.

Suppose now that only output $y(t_k)$ is available for measurement at the periodic measuring instants

$$t_k = kT_c + \delta T_c \qquad (3.13)$$

for some $0<\delta<1$. The timing diagram of the measurement and control instants is depicted in Fig.3.1. The main issue of this section is to examine the effects of the delay factor δ in (3.13) on the state prediction problem. Since the measurement and control instants are not synchronous, the steady state prediction equations must be solved across the time instants kT_c at which the control changes.

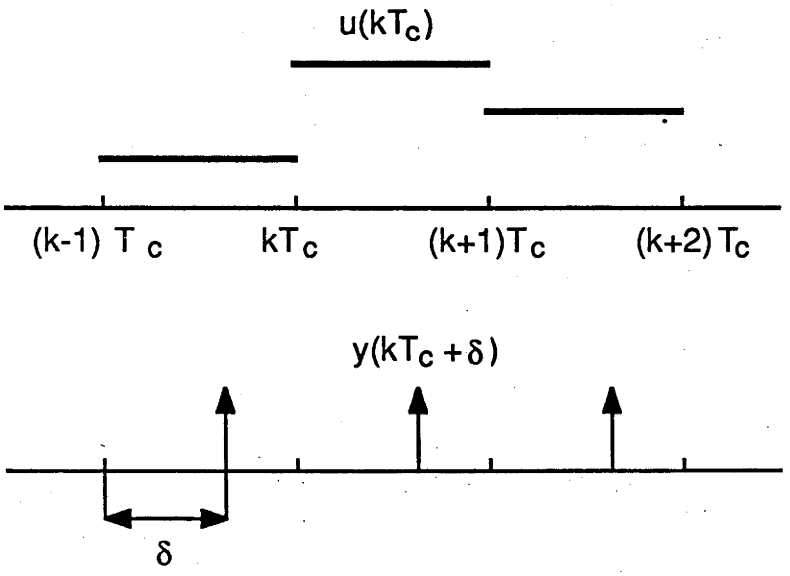


Fig.3.1 Timing diagram of non-synchronous control and output sampling

For simplicity, we assume that the period of the WSCS process disturbance $\omega(t)$ and the measurement noise $\eta(t)$ are equal to the sampling period of the compensator; that is, in (3.11a-b) $T_\omega=T_\eta=T_c$. The discrete equivalent description of

the continuous-time model (3.1a-b) at the measurement instants $t=kT_c+\delta$ and at the control instants $t=kT_c$ can be written as

$$x((k+1)T_c) = \Phi_{1-\delta}x((k+\delta)T_c) + \Gamma_{1-\delta}u(kT_c) + \tilde{\omega}_{1-\delta}((k+\delta)T_c) \quad (3.14a)$$

$$x((k+1+\delta)T_c) = \Phi_{\delta}x((k+1)T_c) + \Gamma_{\delta}u((k+1)T_c) + \omega_{\delta}((k+1)T_c) \quad (3.14b)$$

where the matrices Φ_δ and Γ_δ are given by

$$\Phi_\delta = e^{\delta A T_c} \quad (3.15a)$$

$$\Gamma_\delta = \int_0^{\delta T_c} e^{A(\delta T_c-\sigma)}B d\sigma \quad (3.15b)$$

and the matrices $\Phi_{1-\delta}$ and $\Gamma_{1-\delta}$ are also given by (3.15a) by replacing δ with $1-\delta$. The discrete processes $\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\}$ and $\{\omega_\delta((k+1)T_c)\}$ are respectively given by

$$\tilde{\omega}_{1-\delta}((k+\delta)T_c) = \int_0^{(1-\delta)T_c} e^{A((1-\delta)T_c-\sigma)}\omega((k+\delta)T_c+\sigma)d\sigma \quad (3.16a)$$

$$\omega_\delta((k+1)T_c) = \int_0^{\delta T_c} e^{A(\delta T_c-\sigma)}\omega((k+1)T_c+\sigma)d\sigma \quad (3.16b)$$

The output equation at the measurement instants is governed by

$$y((k+\delta)T_c) = Lx((k+\delta)T_c) + \eta((k+\delta)T_c) \quad (3.17a)$$

where L=C and the covariance of the discrete process $\{\eta((k+\delta)T_c\}$ is given by

$$\Lambda_\delta = \frac{\Lambda_c(\delta T_c)}{T_c} \quad (3.17b)$$

where $\Lambda_c(\delta T_c)$ is defined in (3.11b) for $t=\delta T_c$.

From discrete model (3.14a-b), the one-step-ahead prediction error equations can be derived by extending the standard prediction error equation (3.7d). The result is stated in the following lemma.

Lemma 3.2 Consider the discrete-time model (3.14a-b). Consider also the one-step-ahead predictor equations

$$\hat{x}((k+1)T_c|(k+\delta)T_c) = \Phi_{1-\delta}\hat{x}((k+\delta)T_c|y((k-1+\delta)T_c)) + \Gamma_{1-\delta}u(kT_c) + K_\delta(y((k+\delta)T_c) - L\hat{x}((k+\delta)T_c|y((k-1+\delta)T_c)) \quad (3.18a)$$

$$\hat{x}((k+1+\delta)T_c|y((k+\delta)T_c)) = \Phi_\delta \hat{x}((k+1)T_c|y((k+\delta)T_c)) + \Gamma_\delta u((k+1)T_c) \quad (3.18b)$$

where Φ_δ, Γ_δ are defined in (3.15a-b) and the Kalman filter gain K_δ is to be determined for a certain delay factor δ. Define the prediction error

$$\varepsilon(kT_c) = x(kT_c) - \hat{x}(kT_c|(k-1+\delta)T_c) \quad (3.19)$$

Then, the prediction error equations are governed by

$$\varepsilon((k+1)T_c) = (\Phi_{1-\delta} - K_\delta L)\varepsilon((k+\delta)T_c) - K_\delta \eta((k+\delta)T_c) + \tilde{\omega}_{1-\delta}((k+\delta)T_c) \quad (3.20a)$$

$$\varepsilon((k+1-\delta)T_c) = (\Phi_1 - \Phi_\delta K_\delta L)\varepsilon((k+\delta)T_c) + v_\delta((k+1)T_c) - \Phi_\delta K_\delta \eta((k+\delta)T_c) \quad (3.20b)$$

where the sequence $\{v((k+1)T_c)\}$ is defined by

$$v((k+1)T_c = \omega_\delta((k+1)T_c + \Phi_\delta \tilde{\omega}_{1-\delta}((k+\delta)T_c) \quad (3.20c)$$

and where the processes $\{\tilde{\omega}_{1-\delta}((k+1)T_c)\}$ and $\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\}$ are defined in (3.16a-b).

Proof: The prediction error equation (3.20a) was derived by subtracting the prediction equation (3.18a) from the model (3.14a). From (3.14b) and (3.18b), we obtain

$$\varepsilon((k+1+\delta)T_c) = \Phi_\delta \varepsilon((k+1)T_c) + \omega_\delta((k+1)T_c) \quad (3.21)$$

Substitute $\varepsilon((k+1)T_c)$ in (3.21) using (3.20a) and make use of (3.20c) to get the prediction error equations (3.20b).

□□□

The covariance matrix V_δ of the zero mean process $\{v_\delta((k+1)T_c)\}$ defined in (3.20c) can be calculated according to

$$V_\delta \triangleq \xi\{v_\delta((k+1)T_c)v_\delta'((k+1)T_c)\} = \Omega_\delta + \Phi_\delta \tilde{\Omega}_{1-\delta}\Phi_\delta' + \Phi_\delta X_\delta + X_\delta \Phi_\delta' \quad (3.22)$$

where the covariance matrices Ω_δ and $\tilde{\Omega}_{1-\delta}$ of the processes $\{\omega_\delta((k+1)T_c)\}$ and $\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\}$ respectively can be obtained from (3.16a-b) as

$$\Omega_\delta \triangleq \xi\{\omega_\delta((k+1)T_c)\omega_\delta'((k+1)T_c)\}$$

$$= \int_0^{\delta T_c}\int_0^{\delta T_c} e^{A(\delta T_c-\sigma)} r_\omega(\sigma,\tau) e^{A'(\delta T_c-\tau)} d\sigma d\tau \qquad (3.23a)$$

$$\tilde{\Omega}_{1-\delta} \triangleq \xi\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\tilde{\omega}'_{1-\delta}((k+\delta)T_c)\}$$

$$= \int_0^{(1-\delta)T_c}\int_0^{(1-\delta)T_c} e^{A((1-\delta)T_c-\sigma)} r_\omega(\sigma+\delta T_c,\tau+\delta T_c) e^{A'((1-\delta)T_c-\tau)} d\sigma d\tau \qquad (3.23b)$$

The covariance matrix X_δ in (3.22) can be derived from (3.16a–b), we obtain

$$X_\delta \triangleq \xi\{\omega_\delta((k+1)T_c)\tilde{\omega}'_{1-\delta}((k+\delta)T_c)\}$$

$$= \int_0^{\delta T_c}\int_0^{(1-\delta)T_c} e^{A(\delta T_c-\sigma)} r_\omega(\sigma,\tau+\delta T_c) e^{A'((1-\delta)T_c-\tau)} d\sigma d\tau \qquad (3.23c)$$

Note that in (3.22) $V_1=V_0=\Omega_1$. This follows since at $\delta=0$ or $\delta=1$ in (3.23c) the covariance matrix $X_\delta=0$. If δ were allowed to be zero or one, then $\delta=0$ (or $\delta=1$) correspond to making the measurement synchronous with the control changes. For $\delta=0$, the prediction is made one period ahead whereas for $\delta=1$ the prediction is done 'instantaneuosly'.

If the continuous disturbance $\omega(t)$ is independent WSS then from (3.11a), $r_\omega(t,s)=0$ unless t=s. In this case, from (3.22) and (3.23c) we have

$$V_\delta = \Omega_1 \text{ and } X_\delta = 0$$

for all $\delta\epsilon(0,1)$. Optimal state prediction when both continuous-time processes $\omega(t)$ and $\eta(t)$ are independent WSS and uncorrelated is therefore achieved with $\delta=1$. In practice, some finite delay between the measurement instant and control change is necessary in order to allow time t_c for the calculations to be performed in which case $t_c=(1-\delta)T_c$. This is the so called computational delay (which will be discussed in chapter 4).

We shall show that for the WSCS process disturbance and measurement noise, a synchronous measurement and control is *not* necessarily optimal.

In the conventional problem, the optimal solution is given by the Kalman filter gain (3.8a) which minimizes the trace of the covariance matrix P of the prediction

error $\epsilon(kT_c)$ as in (3.7d). Define Q_δ to be the covariance matrix of the prediction errors at the control instants

$$Q_\delta \triangleq \xi\{\epsilon(kT_c)\epsilon'(kT_c)\} \tag{3.24}$$

Then, the aim of the non-sysnchronous state prediction is to find a certain Kalman filter gain K_δ which solves the following minimization problem

$$\min_{0<\delta<1} \; tr(Q_\delta) \tag{3.25}$$

The following theorem is useful for solving the non-synchronous prediction problem defined by (3.25).

Theorem 3.1 Consider the minimal continuous-time plant (3.1a-b) where the process disturbance $\omega(t)$ and measurement noise $\eta(t)$ are zero mean WSCS of period $T_\omega=T_\eta=T_c$ with covariances defined in (3.11a-b) where $\Omega_c(\tau)\geq 0$ and $\Lambda_c(\tau)>0$ for all $\tau\in[0,T_c]$ where T_c is the control sampling period. Assume the output $y(t_k)$ is available for measurement at the periodic measuring instants t_k defined in (3.13) for $\delta\in(0,1)$.

Then, for a fixed $\delta\in(0,1)$ the optimal steady state predictor which minimizes the criterion (3.25) is described by (3.18a-b) where the Kalman filter gain K_δ satisfies

$$K_\delta = \Phi_{1+\delta}P_\delta L'(\Lambda_\delta + LP_\delta L')^{-1} \tag{3.26a}$$

where $P_\delta \triangleq \xi\{\epsilon((k+\delta)T_c)\epsilon'((k+\delta)T_c)\}$ where the prediction error $\epsilon((k+\delta)T_c)$ is defined in (3.19), and satisfies the algebraic Riccati equation (ARE)

$$P_\delta = \Phi_1 P_\delta \Phi_1' + V_\delta - \Phi_1 P_\delta L'(\Lambda_\delta + LP_\delta L')^{-1}LP_\delta\Phi_1' \tag{3.26b}$$

where L=C in (3.1b), Φ_δ is given by (3.15a), the covariance matrix V_δ is defined in (3.22) and Λ_δ is given by (3.17b).

Furthermore the covariance matrix Q_δ defined in (3.24) is governed by

$$Q_\delta = \bar{\Phi}_{1-\delta}P_\delta\bar{\Phi}_{1-\delta}' + K_\delta\Lambda_\delta K_\delta' + \tilde{\Omega}_{1-\delta} \tag{3.27a}$$

where $\tilde{\Omega}_{1-\delta}$ is defined by (3.23b) and where

$$\bar{\Phi}_{1-\delta} = \Phi_{1-\delta} - K_\delta L \tag{3.27b}$$

Proof: Define

$$K_{1-\delta} \triangleq \Phi_{1-\delta}^{-1} K_\delta \tag{3.28a}$$

Substitute K_δ in (3.20b) using (3.28a), we obtain

$$\varepsilon((k+1+\delta)T_c) = \Phi_1 \varepsilon((k+\delta)T_c) + v_\delta((k+1)T_c) - \Phi_1 K_{1-\delta}(L\varepsilon((k+\delta)T_c)+\eta((k+\delta)T_c)) \tag{3.28b}$$

Note that for a fixed $\delta \in (0,1)$ the processes $v_\delta((k+1)T_c)$ and $\eta((k+\delta)T_c)$ are WSS. Thus, P_δ which is the covariance matrix of the prediction error at the measurement instants $t_k=(k+\delta)T_c$ can be derived from (3.28b), we get

$$P_\delta = \Phi_1 P_\delta \Phi_1' + V_\delta - \Phi_1 P_\delta K_{1-\delta}(LP_\delta L' + \Lambda_\delta)^{-1} K_{1-\delta}' P_\delta' \Phi_1' \tag{3.28c}$$

It is known [Gelb (1974), Anderson and Moore (1979)] that for a fixed $\delta \in (0,1)$, the quantity $\mathrm{tr}(P_\delta)$ where P_δ is defined by (3.28c) is minimized by the Kalman filter gain $K_{1-\delta}$ which satisfies

$$K_{1-\delta} = \Phi_1 P_\delta L' (LP_\delta L' + \Lambda_\delta)^{-1} \tag{3.28d}$$

Substitution of $K_{1-\delta}$ in (3.28c) using (3.28d) gives (3.26b). From (3.28a) and (3.28d), the gain K_δ in (3.26a) can be obtained. The covariance equation Q_δ in (3.27a) can be derived from the prediction equation (3.20a).

□□□

Example 3.1 Consider the following first order plant

$$\dot{x}(t) = -ax(t) + u(t) + \omega(t)$$

$$y(t) = x(t) + \eta(t)$$

where the measurement noise $\eta(t)$ is zero-mean independent WSS with unity variance and the process disturbance $\omega(t)$ is WSCS of period T_ω=1sec (as defined by the first order representation in example 2.1 of chapter 2) with the following statistics

$$r_\omega(t,s) = e^{-\beta(t,s)} + \frac{(1-e^{-\beta t})(1-e^{-\beta s})(1-e^{-2\beta T_\omega})}{(1-e^{-\beta T_\omega})^2}$$

The processes $\{\omega(t)\}$ and $\{\eta(t)\}$ are uncorrelated (ie. $r_{\omega\eta}(t,s)=0$ for all t and s). The continuous-time model was discretized at the sampling period T_c=1sec.

The corresponding graphs of covariances V_δ and $\tilde{\Omega}_{1-\delta}$ defined in (3.22) and (3.23b) for a=2 and β=1 are illustrated in Fig.3.2a.

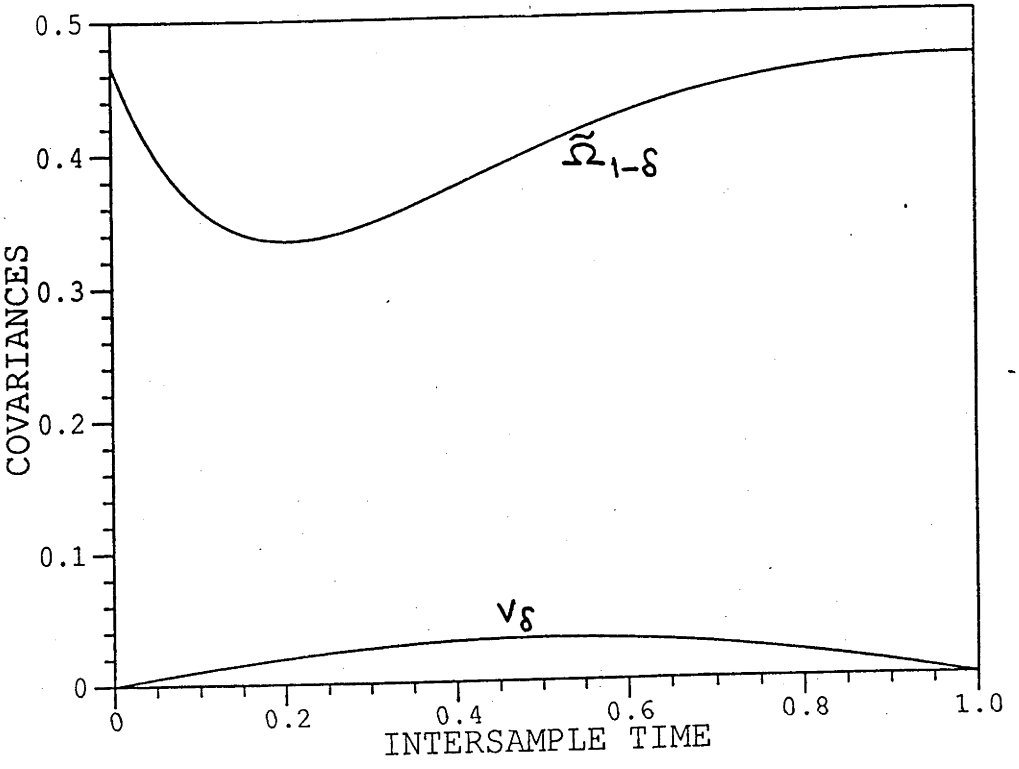


Fig.3.2a Noise covariances V_δ and $\tilde{\Omega}_{1-\delta}$.

The plots of covariance of prediction errors P_δ and Q_δ defined in (3.26b) and (3.24) respectively are depicted in Fig.3.2b. The effects of cyclostationary process $\{\omega(t)\}$ on the state prediction problem can be seen through the prediction error models (3.20a-b) which are driven by WSCS sequences $\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\}$ and $\{v_\delta((k+1)T_c)\}$. It can be seen in Fig.3.2b that the minimum prediction error at the measurement instants (denoted by the minimum of P_δ in Fig.3.2b) does not necessarily produce minimum prediction error at the control instants. As in (3.27a), the optimum selection of the delay factor δ (which minimizes Q_δ) is determined by both P_δ and the WSCS process $\{\tilde{\omega}_{1-\delta}((k+\delta)T_c)\}$. From Fig.3.2b, the optimum delay factor δ depicted by the minimum Q_δ is given by δ^*=0.83.

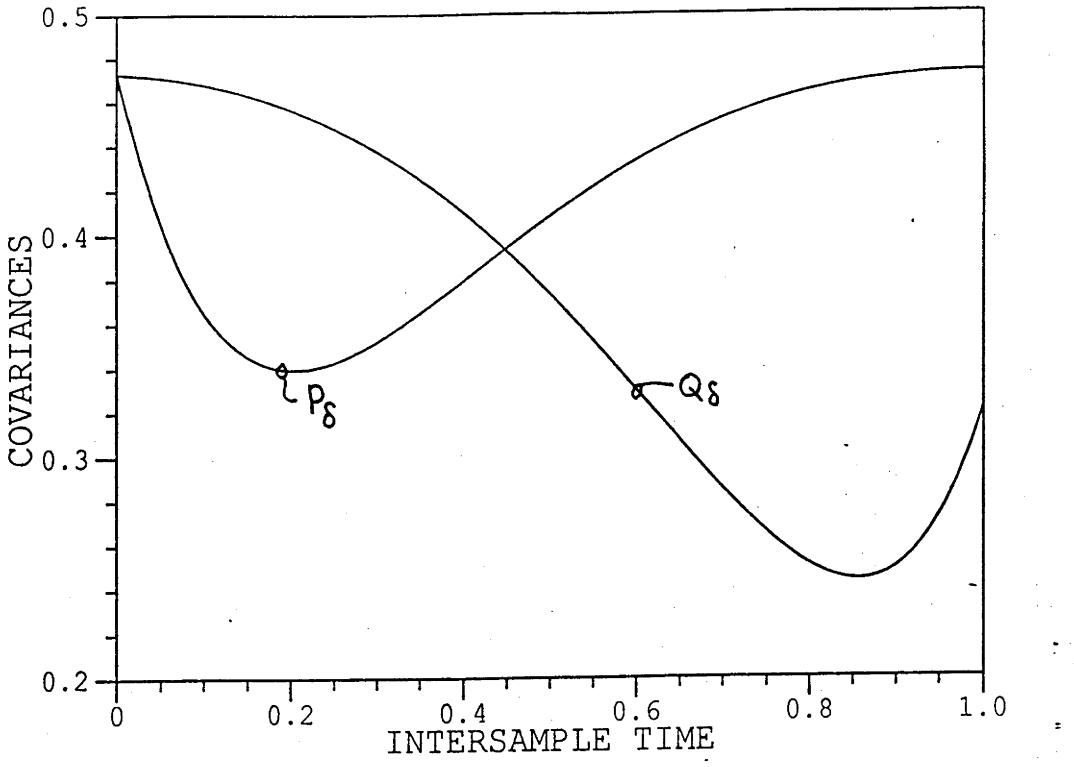


Fig.3.2b Error covariances Q_δ and P_δ.

3.3 MULTIRATE OPTIMAL STATE PREDICTION

Consider the minimal continuous-time system (3.1a-b) where the measurement noise $\eta(t)$ and process disturbance $\omega(t)$ are both WSCS of period T_η and T_ω with known covariances as described in (3.11a-b). Suppose now only the output y(t) is available for measurement at the periodic measuring instant $t_m=jT_m$ and also suppose that the discrete equivalent description of the continuous-time model (3.1a-b) is required at the time instants $t_c=kT_c$ where T_c and T_m are related by (3.10). This circumstance arises when the control signal u(t) has to be updated every T_c seconds.

The timing diagram of measurement and control instants is illustrated in Fig.3.3. From (3.10) and Fig.3.3, the measurement and the control instants are synchronous when $nT_m=kT_c$ or when n=kN where N is defined in (3.10).

In this section, we investigate the consequences of cyclostationary processes $\{\omega(t)\}$ and $\{\eta(t)\}$ on the multirate state prediction problem. Let $y(kT_c-jT_m)$ for all

$j\epsilon[0,N-1]$ be the measurements within a period T_c. Define the data $Y(\alpha,N)$ as

$$Y(\alpha,N) \triangleq \{y(kT_c-(1+\alpha)T_m), y(kT_c-(2+\alpha)T_m), \ldots, y(kT_c-(N+\alpha)T_m)\} \quad (3.29)$$

where N is defined in (3.10) and α is an integer in the interval $0\leq\alpha<N$.

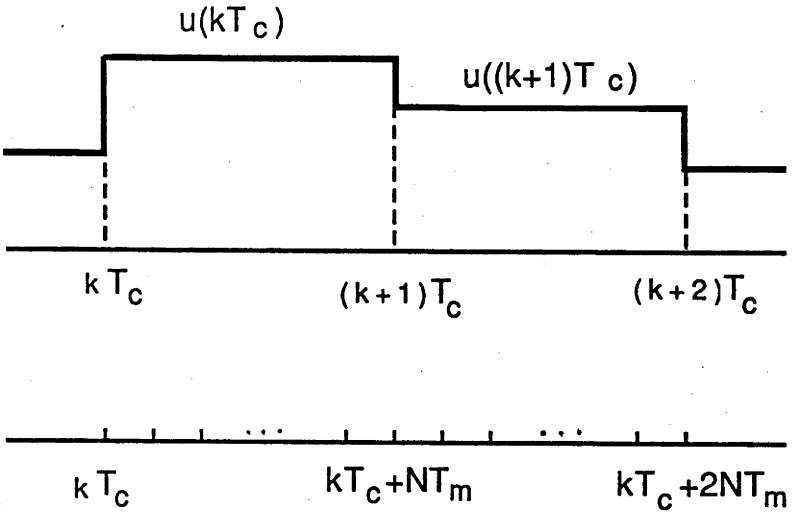


Fig.3.3 Timing diagram of measurement and control instants

The state predictor $\hat{x}(kT|Y(\alpha,N))$ is constructed using the data $Y(\alpha,N)$ in (3.29). We shall show that the optimal state prediction at the instants $t_c=kT_c$ is determined by the selection of the shift factor α when (either one or both of) the processes $\{\omega(t)\}$ and $\{\eta(t)\}$ are WSCS. This is similar to the selection of the delay factor δ in the non-synchronous prediction problem defined in (3.25).

The discrete equivalent description of the continuous-time model (3.1a-b) for a sampling period T_c is given by (3.3a-b). At the measurement instants $t_m=nT_m$, the discrete-time equivalent description of the continuous-time plant (3.1a-b) can be described as follows. For a fixed $\alpha\epsilon[0,N)$,

$$x(kT_c-(i+\alpha+1)T_m) = \Phi_m x(kT_c-(i-\alpha+2)T_m) + \Gamma_m u(kT_c-(i+\alpha+2)T_m) + \omega_m(kT_c-(i+\alpha+2)T_m) \quad (3.30a)$$

$$y(kT_c-(i+\alpha+2)T_m) = Lx(kT_c-(i+\alpha+2)T_m) + \eta(kT_c-(i+\alpha+2)T_m) \quad (3.30b)$$

where the integer $i=0,\pm1,\pm2,\ldots$, and the matrices Φ_m, Γ_m and L are given by

$$\Phi_m = e^{AT_m} \tag{3.31a}$$

$$\Gamma_m = \int_0^{T_m} e^{A(T_m-\sigma)} B d\sigma \tag{3.31b}$$

$$L = C \tag{3.31c}$$

The covariance of the discrete processes $\{\omega_m(kT_c-(i+\alpha+2)T_m)\}$ and $\{\eta(kT_c-(i+\alpha+2)T_m)\}$ are given by

$$\Omega_{i\alpha} = \int_0^{T_m}\int_0^{T_m} e^{A(T_m-\sigma)} r_\omega(kT_c-\gamma T_m+\sigma, kT_c-\gamma T_m-\tau) e^{A'(T_m-\tau)} d\sigma d\tau \tag{3.32a}$$

$$\Lambda_{i\alpha} = \left.\frac{\Lambda_c(t)}{T_c}\right|_{t=kT_c-\gamma T_m} \tag{3.32b}$$

where $\gamma=i+\alpha+2$ and $r_\omega(t,s)$ is defined by (3.11a) and where $\Lambda_c(t)$ is defined in (3.11b). Note that the covariances $\Omega_{i\alpha}$ and $\Lambda_{i\alpha}$ are both determined by i and α due to the cyclostationarity of the processes $\{\omega(t)\}$ and $\{\eta(t)\}$. The sequence $\{u(kT_c)\}$ is assumed to be updated every T_c seconds. Therefore, the sequence $\{u(kT_c-(i+\alpha+2)T_m)\}$ in (3.30a) can be written as

$$u(kT_c-(i+\alpha+2)T_m) = \begin{cases} u((k-2)T_c) & \text{for } N\leq(i+\alpha)<2N \\ u((k-1)T_c) & \text{for } 0\leq(i+\alpha)<N \end{cases} \tag{3.33}$$

The update of the sequence $\{u(kT_c)\}$ is illustrated in Fig.3.3.

If the one-step-ahead predictor were to be constructed at the measurement instant $t_m=nT_m$, for fixed i and α it would be given by

$$\hat{x}(kT_c-(i+\alpha+1)T_m) = \Phi_m\hat{x}(kT_c-(i+\alpha+2)T_m) + \Gamma_m u(kT_c-(i+\alpha+2)T_m) + K_i(y(kT_c-(i+\alpha+2)T_m)-L\hat{x}(kT_c-(i+\alpha+2)T_m)) \tag{3.34a}$$

where

$$\hat{x}(kT_c-(i+\alpha+1)T_m) \triangleq \hat{x}(kT_c-(i+\alpha+1)T_m | y(kT_c-(i+\alpha+2)T_m)) \tag{3.34b}$$

is the one-step-ahead predictor, and K_i is the Kalman filter gain which is to be determined. Define the prediction error $\varepsilon(kT_c-(i+\alpha+1)T_m)$ as

$$\varepsilon(kT_c-(i+\alpha+1)T_m) \triangleq x(kT_c-(i+\alpha+1)T_m) - \hat{x}(kT_c-(i+\alpha+1)T_m) \tag{3.35a}$$

From (3.30a) and (3.34a), we obtain the prediction error equation

$$\varepsilon(kT_c-(i+\alpha+1)T_m) = \Phi_m\varepsilon(kT_c-(i+\alpha+2)T_m) + \omega_m(kT_c-(i+\alpha+2)T_m) - K_i(y(kT_c-(i+\alpha+2)T_m) - L\hat{x}(kT_c-(i+\alpha+2)T_m)) \quad (3.35b)$$

The optimal Kalman filter gain K_i is given by

$$K_i = \Phi_m P_i L' (\Lambda_{i\alpha}+L' P_i L)^{-1} \quad (3.36a)$$

where for fixed i and α the covariance $\Lambda_{i\alpha}$ is defined in (3.32b) and the covariance matrix P_i satisfies the ARE

$$P_i \triangleq \xi\{\varepsilon(kT_c-(i+\alpha+1)T_m)\varepsilon'(kT_c-(i+\alpha+1)T_m)\}$$
$$= \Phi_m P_i \Phi_m' + \Omega_{i\alpha} - \Phi_m P_i L' (\Lambda_{i\alpha}+LP_i L')^{-1} LP_i \Phi_m' \quad (3.36b)$$

where Φ_m and L are defined by (3.31a) and (3.31c) and where the the covariance $\Omega_{i\alpha}$ is defined in (3.32a).

Using the one-step-ahead prediction equation (3.34a), the required predictor $\hat{x}(kT_c|Y(\alpha,N))$ can be constructed. The result is stated in the following lemma.

Lemma 3.3 Consider the minimal continuous-time model (3.1a-b) where the measurement and control sampling are related by the multirate sampling (3.10) for a certain positive integer N. Assume the process disturbance $\omega(t)$ and the measurement noise $\eta(t)$ are both WSCS with periods T_ω and T_η, uncorrelated and having covariances $\Omega_c(t)$ and $\Lambda_c(t)$ defined in (3.11a-b). Assume $T_\omega=T_\eta=T_c$. Consider as well the one-step-ahead prediction equation (3.34a) with the Kalman filter gain K_i and the covariance P_i defined by (3.36a-b).

Then at a fixed $\alpha\in[0,N]$, the predicted state $\hat{x}(kT_c)$ (which is a simplified notation of $\hat{x}(kT_c|Y(\alpha,N))$) based on the N measurement $Y(\alpha,N)$ in (3.29) is given by

$$\hat{x}(kT_c-(1+\alpha)T_m) = \Phi\hat{x}((k-1)T_c-(1+\alpha)T_m) + \sum_{j=N-\alpha-1}^{N-1} \Phi_m^j\Gamma_m u((k-2)T_c) + \sum_{j=0}^{N-\alpha-2} \Phi_m^j\Gamma_m u((k-1)T_c) +$$

$$[\Phi_m^{N-1}K_1 \ \ldots \ \Phi_m K_{N-1} \ K_N] \begin{bmatrix} y((k-1)T_c-(1+\alpha)T_m)-\hat{y}((k-1)T_c-(1+\alpha)T_m) \\ y((k-1)T_c-\alpha T_m)-\hat{y}((k-1)T_c-\alpha T_m) \\ \vdots \\ y((k-1)T_c-(2+\alpha-N)T_m)-\hat{y}((k-1)T_c-(2+\alpha-N)T_m) \end{bmatrix} \tag{3.37a}$$

$$\hat{x}(kT_c) = \Phi_m^{1+\alpha}\hat{x}(kT_c-(1+\alpha)T_m) + \sum_{j=0}^{\alpha} \Phi_m^j \Gamma_m u((k-1)T_c) \tag{3.37b}$$

where the matrices Φ_m and Γ_m are given by (3.31a-b) and the Kalman filter gain K_i for all $i\epsilon[1,N]$ are given by (3.36a).

Furthermore, the prediction error $\epsilon(kT_c)$ defined by (3.35a) is governed by

$$\begin{aligned} \epsilon(kT_c-(1+\alpha)T_m) = (\Phi - \sum_{j=0}^{N-1} \Phi_m^{N-1-j}K_{j+1}L)\epsilon((k-1)T_c-(1+\alpha)T_m)) + \\ \sum_{j=0}^{N-1} \Phi_m^{N-1-j}\omega_m((k-1)T_c-(1+\alpha-j)T_m) + \\ \sum_{j=0}^{N-1} \Phi_m^{N-1-j}K_{j+1}\eta_m((k-1)T_c-(1+\alpha-j)T_m) \end{aligned} \tag{3.38a}$$

$$\epsilon(kT_c) = \Phi_m^{1+\alpha}\epsilon(kT_c-(1+\alpha)T_m) + \sum_{j=0}^{\alpha} \Phi_m^{\alpha-j}\omega_m(kT_c-(1+\alpha-j)T_m) \tag{3.38b}$$

where the processes $\{\omega_m(.)\}$ and $\{\eta_m(.)\}$ are the discrete version of the continuous-time processes $\{\omega(t)\}$ and $\{\eta(t)\}$, and are defined in (3.30a-b).

Proof: The predictor (3.37a) can be derived by manipulating N of the one-step-ahead predictor (3.34a) and make use the relation between Φ in (3.3c) and Φ_m in (3.31a); that is

$$\Phi = \Phi_m^N$$

The predictor $\hat{x}(kT_c)$ in (3.37b) is the propagation of the predictor $\hat{x}(kT_c-(1+\alpha)T_m)$ without the influence of the Kalman filter gains K_i. From the discrete-time model (3.30a), we obtain the following models

$$x(kT_c-(1+\alpha)T_m) = \Phi x((k-1)T_c-(1+\alpha)T_m) +$$

$$\sum_{j=N-\alpha-1}^{N-1} \Phi_m{}^j\Gamma_m u((k-2)T_c) + \sum_{j=0}^{N-\alpha-2} \Phi_m{}^j\Gamma_m u((k-1)T_c) +$$

$$\sum_{j=0}^{N-1} \Phi_m{}^j\omega_m((k-1)T_c-(1+\alpha-j)T_m) \qquad (3.39a)$$

$$x(kT_c) = \Phi_m{}^{1+\alpha}x(kT_c-(1+\alpha)T_m) +$$

$$\sum_{j=0}^{\alpha} \Phi_m{}^{\alpha-j}\omega_m(kT_c-(1+\alpha-j)T_m) \qquad (3.39b)$$

The prediction error equation (3.38a) can be derived by subtracting the prediction equation (3.37a) from the model (3.39a). The prediction error equation (3.38b) can be obtained by subtracting the prediction equation (3.37b) from the model (3.39b).

□□□

Note that the prediction errors $\varepsilon(kT_c-(1+\alpha)T_m)$ and $\varepsilon(kT_c)$ in (3.38a-b) are both affected by the shift factor α which determines the data $Y(\alpha,N)$ in (3.29). The multirate state prediction problem can then be formulated as the problem of finding the shift factor α assuming the Kalman filter gains K_i in (3.36a) are given for all i in the interval $1\leq i\leq N$ such that the prediction error at the control instants $t=kT_c$ is minimized; more precisely

$$\min_{0\leq\alpha<N} \; tr[\xi\{\varepsilon(kT_c)\varepsilon'(kT_c)\}] \qquad (3.40)$$

where $\varepsilon(kT_c)$ is given by (3.38b). When the processes $\omega(t)$ and $\eta(t)$ are both WSS, the covariances $\Omega_{i\alpha}$ in (3.32a) and $\Lambda_{i\alpha}$ in (3.32b) are both constants for all i and α. Therefore, from (3.36a-b) it can be deduced that K_i are equal for all i Consequently, the optimal selection of the shift factor α in (3.40) is zero.

The following theorem establishes the multirate state predictor under the influence of cyclostationary disturbances.

Theorem 3.2 Consider the minimal continuous-time model (3.1a-b). Suppose the plant output $y(t_m)$ is available at the instants $t_m=nT_m$ and suppose the plant input $u(t)$ (which is assumed to be an output of a digital compensator) changes every T_c seconds where the sampling periods T_c and T_m are related by the multirate sampling (3.10) for a certain positive integer N. Assume the process disturbance $\omega(t)$ and the measurement noise $\eta(t)$ are both WSCS with periods T_ω and T_η, uncorrelated and having covariances $\Omega_c(t)$ and $\Lambda_c(t)$ defined in (3.11a-b), and also assume $T_\omega=T_\eta=T_c$. Consider as well the discrete-time equivalent description (3.30a-b).

Then, for a fixed $\alpha\epsilon[0,N]$ the multirate state predictor is governed by (3.37a-b) and the corresponding prediction error equation is given by (3.38a-b) where the Kalman filter gains K_i and the covariance matrix P_i for $i\epsilon[1,N]$ are given by (3.36a-b). Moreover, the optimal multirate state predictor is the one which minimizes the criterion (3.40).

□□□

Example 3.2 Consider again the scalar model used in example 3.1; that is,

$$\dot{x}(t) = -ax(t) + u(t) + \omega(t)$$

$$y(t) = x(t) + \eta(t)$$

where the measurement $\eta(t)$ is zero-mean independent WSS with unity variance and the process disturbance $\omega(t)$ is WSCS with period T_ω=1sec., and where $\omega(t)$ and $\eta(t)$ are uncorrelated. The control signal $u(t)$ (produced by a digital computer) changes every T_c=1sec. The second-order statistics of the process $\{\omega(t)\}$ is given by

$$r_\omega(t,s) = e^{-\beta(t,s)} + \frac{(1-e^{-\beta t})(1-e^{-\beta s})(1-e^{-2\beta T_\omega})}{(1-e^{-\beta T_\omega})^2}$$

for all t and s in the interval $[0,T_\omega]$. Within one period T_c, the plant output $y(t)$ is measured 10 times. Therefore, N=10 in (3.10) and T_m=0.1sec. For each fixed $i\epsilon[1,N]$ and for a=2 and β=1, the Kalman filter gain K_i can be computed by finding the steady-state solution of (3.36a-b). By means of lemma 3.3 and theorem 3.2, the

covariance of the prediction errors $\epsilon(kT_c)$ and $\epsilon(kT_c-(1+\alpha)T_m)$ defined in (3.38a-b) can be computed for each $\alpha\epsilon[0,N]$. The result is presented in table 3.1. It can be seen in table 3.1 that the minimum prediction error at the measurement instant $t_m=kT_c-6T_m$ (ie. $\alpha=5$) does not necessarily produce minimum prediction error at the control instants. The optimal shift factor α which minimizes the criterion (3.40) is given by $\alpha^*=0.2$.

α	$\xi\{\epsilon^2(kT_c-(1+\alpha)T_m)\}$	$\xi\{\epsilon^2(kT_c)\}$
9	0.3951	0.3961
8	0.3679	0.3867
7	0.2667	0.3672
6	0.2331	0.3238
5	0.1950	0.2661
4	0.2461	0.2626
3	0.2718	0.2215
2	0.3234	0.1943
1	0.3372	0.2121
0	0.3468	0.2347

Table 3.1 The influence of the shift factor α on the state prediction errors

3.4 CONCLUSIONS

In this chapter, the consequences of cyclostationary process disturbance and measurement noise on the state prediction problem have been investigated. In the single rate case (ie. when the measurement and the control sampling rate are equal), it has been shown that the optimal state prediction is not generally achieved by synchronous control and measurement sampling. We have shown that under the influence of cyclostationary disturbances the optimal Kalman filter gain K_δ is determined by the delay factor δ.

In the multirate case when the measurement sampling rate is faster than the control sampling rate (ie. $T_c=NT_m$ for a positive integer N), it has been shown that due to the cyclostationarity of the process disturbance and measurement noise, the

optimal prediction is determined by the selection of the shift factor α.

Both the non-synchronous and the multirate state prediction problems were solved by means of the standard algebraic Riccati technique. When the process disturbance and the measurement noise are both wide-sense stationary, the non-synchronous and the multirate state prediction problems become the conventional single-rate and multirate state prediction problems. When the measurement and control sampling rates are equal, the optimal delay factor δ which satisfies the criterion (3.25) is given by $\delta^*=1$ which means the measurement and control sampling are synchronous and the state prediction $\hat{x}(kT_c)$ is derived 'instantaneously' from the measurement $y(kT_c)$. This is unrealistic. When the measurement sampling rate is faster than the control sampling rate (ie. $T_c=nT_m$), the optimal shift factor α which satisfies the criterion (3.40) is given by $\alpha^*=0$ which means the state prediction $\hat{x}(kT_c)$ is derived from the measurement $y(kT_c-T_m)$. In fact this is the method used in the conventional multirate state prediction design.

CHAPTER 4

DISCRETE LINEAR QUADRATIC REGULATION OF CONTINUOUS-TIME SYSTEMS

4.1 INTRODUCTION

It has been shown in chapter 2 that digital control implementations produce cyclostationary continuous-time responses. This implies the intersample variance may vary significantly from that monitored at the controlling instants. Cyclostationary processes have long been an interest and concern in communications theory [Franks (1969), Gardner and Franks (1975), Ogura (1971)] but the consequences for digital control and estimation have received little attention. Continuous-time performance of discrete minimum variance regulators has been investigated in [de Souza and Goodwin (1984)].

In this chapter, we examine the consequences of discrete linear quadratic regulation of continuous-time time-invariant systems. The conventional discrete linear quadratic regulators are designed completely on the knowledge of the discrete data rensponse. For example, minimum variance regulators [Åström (1970)] are purely derived from the plant output observed at discrete instants. The adaptive version of these regulators [Åström (1977)] are also designed in this way. It has been outlined in [de Souza and Goodwin (1984)] that the classical minimum variance regulators which are the stochastic versions of the deterministic 'dead-beat' controllers often produce (undesirable) high intersample variance. We shall see that the high intersample variance can be reduced by incorporating the intersample behavior in the performance criterion to be optimized.

In section 4.2, we briefly review the ideal (ie. infinite precision) discrete linear

quadratic Gaussian (LQG) regulators. (The finite precision LQG regulators will be discussed in chapters 5 and 6). There are two main points that we discuss in this section. First we review the certainty equivalence principle which allows the estimator and the controller to be designed separately. Then, we briefly discuss the necessary and sufficient conditions for the existence of the stabilizing solutions. In section 4.3, a quadratic cost function is introduced for the purpose of improving the intersample behavior of a discrete regulator when implemented on a continuous plant. Optimization of this performance index results in what we term the *minimax quadratic regulator*. We show that the minimax quadratic regulation problem can be reformulated as a standard LQG problem. A particular example of a minimax quadratic regulator is the minimax output variance regulator which is investigated in section 4.4. The performance of the minimax variance regulator is compared to other methods of output regulation. We demonstrate that the ill conditioning often associated with minimum variance regulation is due in part to the cyclostationarity nature of the controlled variable. It is shown that the proposed minimax solution offers a significant improvement over the standard methods. At the end of section 4.4, we present some illustrative examples.

4.2 LINEAR QUADRATIC GAUSSIAN REGULATOR DESIGN

In this section, we briefly review the ideal linear quadratic Gaussian (LQG) regulator problem and the optimal compensator that results. We wish to design a discrete–time compensator for a continuous–time system and the control signal will be piece–wise constant. We assume that the control sequence $\{u(kT_c)\}$ is updated at the rate T_c^{-1}.

Consider a minimal continuous–time plant described by

$$\dot{x}(t) = Ax(t) + Bu(t) + \omega(t) \tag{4.1a}$$

$$y(t) = Cx(t) + \tilde{\eta}(t) \tag{4.1b}$$

where the state, the input and the output are respectively

$$x(t)\epsilon R^{n_x} \quad ; \quad u(t)\epsilon R^{n_u} \quad ; \quad y(t)\epsilon R^{n_y}$$

and the system matrix A is $(n_x x n_x)$, the input matrix B is $(n_x x n_x)$ and the output matrix C is $(n_y x n_x)$. The process disturbance $\omega(t)$ and the measurement noise $\eta(t)$ are both zero mean *wide-sense stationary* (WSS), are uncorrelated and having covariances

$$\xi\{\begin{bmatrix}\omega(t)\\ \tilde{\eta}(t)\end{bmatrix}[\omega'(t)\ \tilde{\eta}'(t)]\} = \begin{bmatrix}\Omega_c & 0\\ 0 & \Lambda_c\end{bmatrix} \tag{4.2a}$$

We assume that the output y(t) in (4.1b) is passed through an (ideal) pre-filter which passes the term Cx(t) unchanged and only filter the wide-band noise $\tilde{\eta}(t)$. The pre-filtered noise is denoted by $\eta(t)$. The control signal u(t) in (4.1a) is piece-wise constant of the form

$$u(t) = \sum_{k=-\infty}^{\infty} u(kT_c)p(t-kT_c) \tag{4.2b}$$

where

$$p(t) = \begin{cases} 1 & \text{for } t\epsilon[0,T_c) \\ 0 & \text{otherwise} \end{cases} \tag{4.2c}$$

where T_c is a certain sampling period. For the continuous-time LQG problem, the quadratic performance index which will be minimized, can be written as follows

$$J_c = \xi\{\int_{t_0}^{t_1}[x'(t)Q_c x(t) + u'(t)R_c u(t)]dt + x'(t_1)Q_0 x(t_1)\} \tag{4.3}$$

for some weighting matrices $Q_c \geq 0$, $Q_0 \geq 0$ and $R_c > 0$.

The equivalent discrete-time description of the continuous-time model (4.1a-b) for a sampling period T_c is given by

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \tag{4.4a}$$

$$y(kT_c) = Lx(kT_c) + \eta(kT_c) \tag{4.4b}$$

where $\{\eta(kT_c)\}$ is the sampled version of the pre-filtered noise $\eta(t)$, and the matrices Φ, Γ and L and the matrices A, B and C in (4.1a-b) are related by

$$\Phi = e^{AT_c} \tag{4.4c}$$

$$\Gamma = \int_0^{T_c} e^{A(T_c-\sigma)} \tag{4.4d}$$

$$L = C \tag{4.4e}$$

The covariance of the discrete processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$ are given by

$$\Omega \triangleq \xi\{\omega(kT_c)\omega'(kT_c)\}$$

$$= \int_0^{T_c} e^{A(T_c-\sigma)}\Omega_c e^{A'(T_c-\sigma)}\,d\sigma \tag{4.5a}$$

$$\Lambda \triangleq \xi\{\eta(kT_c)\eta'(kT_c)\}$$

$$= \frac{\Lambda_c}{T_c} \tag{4.5b}$$

$$\xi\{\omega(kT_c)\eta'(kT_c) = 0$$

where Ω_c and Λ_c are defined in (4.2a). The discrete-time version J_d of the performance index J_c in (4.3) for a sampling period T_c satisfies

$$\lim_{T_c\to\infty} J_d = J_c$$

where the quadratic cost J_d is given by

$$J_d = \xi\{\sum_{k=-n}^{m-1}[x'(kT_c)Q_d x(kT_c) + 2x'(kT_c)M_d u(kT_c) + u'(kT_c)R_d u(kT_c)] + x'(mT_c)Q_0 x(mT_c)\} \tag{4.6a}$$

where nT_c and mT_c correspond to t_0 and t_1 in (4.3). The weighting matrices Q_d, M_d and R_d are related to the weighting matrices Q_c and R_c by

$$Q_d = \int_0^{T_c} \Phi'(\sigma)Q_c\Phi(\sigma)\,d\sigma \tag{4.6b}$$

$$M_d = \int_0^{T_c} \Phi'(\sigma)Q_c\Gamma(\sigma)\,d\sigma\,. \tag{4.6c}$$

$$R_d = R_c T_c + \int_0^{T_c} \Gamma'(\sigma)Q_c\Gamma(\sigma)\,d\sigma \tag{4.6d}$$

where

$$\Phi(\tau) = e^{A\tau} \tag{4.6e}$$

$$\Gamma(\tau) = \int_0^{\tau} \Phi(\sigma)\,d\sigma \tag{4.6f}$$

The optimal compensator which minimizes the quadratic index (4.6a) subject to the discrete plant (4.4a-b) is governed by

$$\hat{x}((k+1)T_c) = \Phi\hat{x}(kT_c) + \Gamma u(kT_c) + K(kT_c)(y(kT_c) - L\hat{x}(kT_c)) \tag{4.7a}$$

$$u(kT_c) = -G(kT_c)\hat{x}(kT_c) \tag{4.7b}$$

where $\hat{x}(kT_c)$ (which is a simplified notation for $\hat{x}(kT_c|y((k)T_c))$ is the estimated state vector and where the $(n_x x n_y)$ matrix $K(kT_c)$ and the $(n_u x n_x)$ matrix $G(kT_c)$ are respectively the Kalman filter and the controller gains. Notice that in (4.7b) the control $u(kT_c)$ depends on past values of the plant output $y(mT_c)$ up to and including m=k [Sage (1968), Åström and Wittenmark (1984)]. In practice, this type of compensator is not directly feasible for implementation since a certain time must be allowed for the controller to compute $u(kT_c)$ from the past outputs. In other words, some computational delay must be included in the design. The design which allows one full sample period for the computation of the control signal $u(kT_c)$ based on past values of outputs $y(mT_c)$ up to and included m=k-1 is presented in [Kwakernaak and Sivan (1972)]. But this idea would introduce inefficiency if the required computation time is less than one full sample period which means the control signal $u(kT_c)$ will be available before it is actually used. This inefficiency can be avoided by introducing a method of skewing the sample time of the plant output with respect to the rest of the compensator as shown in Fig.4.1. The details of this method can be found in [Kwakernaak and Sivan (1972), Åström and Wittenmark (1984)]. Henceforth, for simplicity we assume no skewing. Note that the delay factor δ that we discussed in the previous chapter (see Fig.3.1) is not allowed to be less than the computational delay.

Now define the prediction error $\epsilon(kT_c)$ as the difference between the true state $x(kT_c)$ and the estimated state $\hat{x}(kT_c)$ as follows.

$$\varepsilon(kT_c) \triangleq x(kT_c) - \hat{x}(kT_c) \tag{4.8}$$

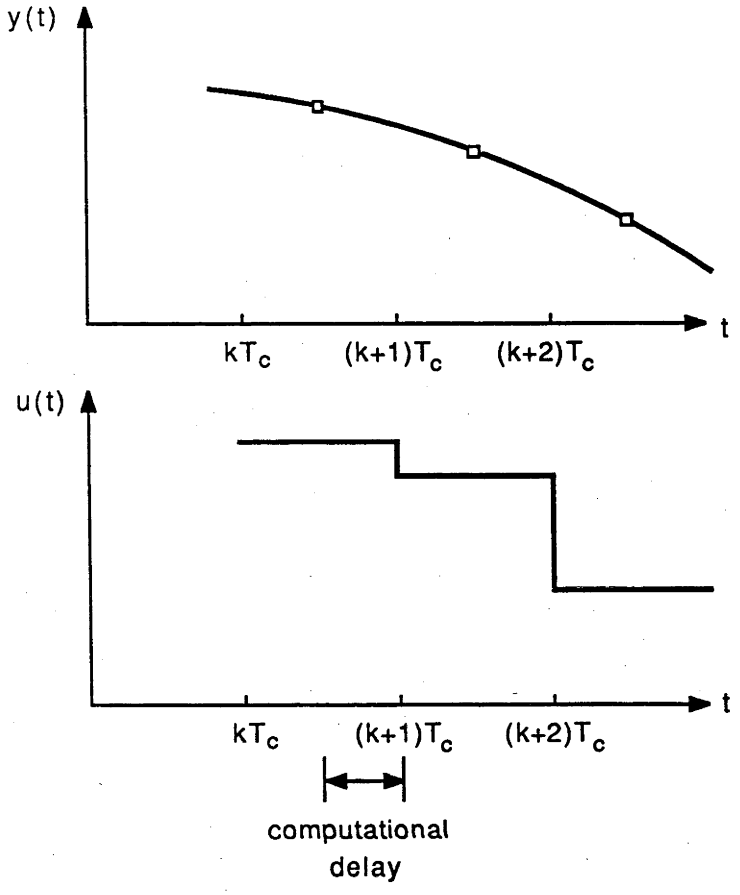


Fig.4.1 The computational delay requirement

From (4.4a-b) and (4.7a-b), the augmented model which includes the prediction error equation can be written as follows

$$\begin{bmatrix} \varepsilon((k+1)T_c) \\ \hat{x}((k+1)T_c) \end{bmatrix} = \begin{bmatrix} \Phi-K(kT_c)L & 0 \\ K(kT_c)L & \Phi-\Gamma G(kT_c) \end{bmatrix} \begin{bmatrix} \varepsilon(kT_c) \\ \hat{x}(kT_c) \end{bmatrix} + \begin{bmatrix} I & -K(kT_c) \\ 0 & K(kT_c) \end{bmatrix} \begin{bmatrix} \omega(kT_c) \\ \eta(kT_c) \end{bmatrix} \tag{4.9}$$

The LQG control problem is then defined as the problem of designing a linear compensator (4.7a-b) which minimizes the quadratic performance index (4.6a) subject to the discrete system (4.4a-b).

The optimal solution to this problem satisfies the *separation theorem* (also known as *certainty equivalent principle*) which implies that the optimal LQG control

problem can be separated into two parts: the state estimator problem (that is, find the optimal estimate of the plant state from the observed outputs) and the linear feedback control problem. Hence, the linear control law $u(kT_c)$ in (4.7b) can be designed assuming the plant state $x(kT_c)$ is available for control. This fact has been widely discussed in the literature, in particular it can be found in [Kwakernaak and Sivan (1972), Franklin and Powell (1981), Åström and Wittenmark (1984)]. The result is stated in the following lemma.

Lemma 4.1 Consider the discrete system (4.4a–b) where the processes $\{\omega(kT_c)\}$ and $\{\eta(kT_c)\}$ are zero-mean discrete Gaussian with respective covariance matrices Ω and Λ, and are uncorrelated. Consider also the linear compensator (4.7a–b) and the quadratic performance index (4.6a) and the corresponding parameters (4.6b–f). Define the covariance of the prediction error

$$P(kT_c) \triangleq \xi\{\varepsilon(kT_c)\varepsilon'(kT_c)\}$$

where $\varepsilon(kT_c)$ is defined by (4.8).

Then the solution of the LQG problem is given by the optimal Kalman gain $K(kT_c)$ which minimizes the prediction error measured by the trace of $P(kT_c)$ and by the optimal regulator feedback gain $G(kT_c)$ which minimizes the quadratic index (4.6a).

Furthermore, the gains $K(kT_c)$ and $G(kT_c)$ are given by

$$K(kT_c) = \Phi P(kT_c)L'(LP(kT_c)L' + \Lambda)^{-1} \tag{4.10a}$$

$$\bar{G}(kT_c) = (R_d + \Gamma'\Sigma((k+1)T_c)\Gamma)^{-1}\Gamma'\Sigma((k+1)T_c)\bar{\Phi} \tag{4.10b}$$

where the matrices $P(kT_c)$ and $\Sigma(kT_c)$ are the solutions of the following discrete Riccati difference equations

$$P((k+1)T_c) = (\Phi - K(kT_c)L)P(kT_c)(\Phi - K(kT_c)L)' + \Omega + K(kT_c)\Lambda K(kT_c)' \tag{4.11a}$$

$$\Sigma(kT_c) = (\bar{\Phi} - \Gamma\bar{G}(kT_c))'\Sigma((k+1)T_c)(\bar{\Phi} - \Gamma\bar{G}(kT_c)) + \bar{Q}_d + \bar{G}(kT_c)'R_d\bar{G}(kT_c) \tag{4.11b}$$

where

$$\bar{\Phi} = \Phi - \Gamma R_{\bar{d}}^{-1}M_{\bar{d}}' \tag{4.11c}$$

$$\bar{G}(kT_c) = G(kT_c) - R_{\bar{d}}^{-1}M_{\bar{d}}' \tag{4.11d}$$

$$\bar{Q}_d = Q_d - M_d R_d^{-1} M_d' \qquad (4.11e)$$

□□□

The independence of the Kalman filter gain $K(kT_c)$ from the controller gain $G(T_c)$ or the independence of the covariance matrix $P(kT_c)$ from $\Sigma(kT_c)$ is a consequence of the optimal solution. The separation principle can be shown through the (asymptotic) independence of the prediction error $\varepsilon(kT_c)$ from the estimated state $\hat{x}(kT_c)$ and can be established from the augmented representation (4.9) and lemma 4.1. Define

$$\xi\{\begin{bmatrix}\varepsilon(kT_c)\\ \hat{x}(kT_c)\end{bmatrix}[\varepsilon'(kT_c) \quad \hat{x}'(kT_c)]\} \triangleq \begin{bmatrix}P(kT_c) & P_{12}(kT_c)\\ P_{12}'(kT_c) & P_{22}(kT_c)\end{bmatrix} \qquad (4.12a)$$

where the covariance matrix $P(kT_c)$ is given by (4.11a). From (4.9), we obtain

$$P_{12}((k+1)T_c) = (\Phi - K(kT_c)L)P(kT_c)L'K'(kT_c) + (\Phi - K(kT_c)L)P_{12}(kT_c)(\Phi - \Gamma G(kT_c))' - K(kT_c)\Lambda K'(kT_c) \qquad (4.12b)$$

where the gains $K(kT_c)$ and $G(kT_c)$ are given by (4.10a–b) and the covariance matrix $P(kT_c)$ is given by (4.11a). Substitute $K(kT_c)$ in (4.12b) using (4.10a), results in the homogenous equation

$$P_{12}((k+1)T_c) = (\Phi - K(kT_c)L)P_{12}(kT_c)(\Phi - \Gamma G(kT_c))' \qquad (4.12c)$$

which will asymptotically approach zero provided $\Phi - K(kT_c)L$ and $\Phi - \Gamma G(kT_c)$ are both stable.

Another type of LQG problem is to minimize the limit of the quadratic cost (provided such value exists)

$$J_c = \xi\{\lim_{t_1\to\infty} \frac{1}{2t_1}\int_{-t_1}^{t_1} [x'(t)Q_c x(t) + u'(t)R_c u(t)]dt \qquad (4.13)$$

with $Q_c \geq 0$ and $R_c > 0$. The discrete-time version J_d of the quadratic cost J_c is given by

$$J_d = \xi\{\lim_{m\to\infty} \frac{1}{2m}\sum_{k=-m}^{m} [x'(kT_c)Q_d x(kT_c) + 2x'(kT_c)M_d u(kT_c) + u'(kT_c)R_d u(kT_c)]\} \qquad (4.14)$$

where Q_d, M_d and R_d are defined by (4.6b-d). This type of LQG problem is known as the *infinite horizon problem* or as steady-state problem. The LQG problem which minimizes the performance index (4.6a) is then called the *finite horizon problem.* The infinite time horizons (ie. $m\to\infty$) both for the optimal state estimation and the optimal regulation reflect the steady-state nature of the optimization [Sage (1968), Kwakernaak and Sivan (1972)]. The optimal solutions K and G are now time-invariant and are given by the steady-state solutions (if they exist) of the discrete *algebraic* Riccati equations; that is the time-invariant version of (4.11a-b). Note that the cost J_d in (4.14) is not well defined unless the closed loop system is asymptotically stable (ie. all eigenvalues of $\Phi-KL$ and $\Phi-\Gamma G$ are strictly inside the unit circle). The steady-state solutions P and Σ of (4.11a-b) which give the stabilizing gains K and G are called *stabilizing solutions.* The necessary and sufficient conditions for stability of the closed loop system are stated in the following lemma which is discussed in detail in [Chan et. al (1984)].

Lemma 4.2 [Chan et. al (1984)] Consider the minimal continuous-time system (4.1a-b) and the (infinite horizon) performance index J_c in (4.13) with $Q_c\geq 0$ and $R_c>0$, and also the corresponding discrete-time equivalent description (4.4a-b) and the performance index J_d in (4.14). Consider as well the linear compensator (4.7a-b) where the (time-invariant) Kalman filter and controller gains K and G satisfy

$$K = \Phi P L' (LPL'+\Lambda)^{-1} \tag{4.15a}$$

$$\bar{G} = (R_d+\Gamma'\Sigma\Gamma)^{-1}\Gamma'\Sigma\bar{\Phi} \tag{4.15b}$$

and the matrices P and Σ satisfy the algebraic Riccati equations

$$P = (\Phi-KL)P(\Phi-KL)' + \Omega + K\Lambda K' \tag{4.16a}$$

$$\Sigma = (\bar{\Phi}-\Gamma\bar{G})'\Sigma(\bar{\Phi}-\Gamma\bar{G}) + \bar{Q}_d + \bar{G}'R_d\bar{G} \tag{4.16b}$$

where $\bar{\Phi}$ is defined in (4.11c), $\bar{Q}_d$ is defined in (4.11e) and

$$\bar{G} = G-R_d^{-1}M_d' \tag{4.16c}$$

Suppose the matrices Ω and Q are factorized as follows

$$\Omega \triangleq DD'$$

$$\bar{Q} = \overline{DD'}$$

Then the Kalman filter and the controller gains K and G are stabilizing if and only if

(i) $\{\Phi,D\}$ has no uncontrollable modes on the unit circle and

(ii) $\{\bar{\Phi},\bar{D}\}$ has no unobservable modes on the unit circle.

□□□

4.3 MINIMAX QUADRATIC REGULATION

Consider the minimal continuous-time time-invariant model (4.1a-b) where the control signal u(t) is piece-wise constant defined by (4.2b-c). Assume the process disturbance $\omega(t)$ and measurement noise $\eta(t)$ are zero-mean and uncorrelated WSS with covariances defined by (4.2a). At the controlling instant $t=kT_c$, the equivalent discrete-time description of the plant (4.1a-b) is given by (4.4a-b). The intersample discrete-time realization of the continuous-time plant (4.1a-b) can be written as

$$x((k+\delta)T_c) = \Phi_\delta x(kT_c) + \Gamma_\delta u(kT_c) + \omega_\delta(kT_c) \qquad (4.17a)$$

$$y((k+\delta)T_c) = Lx((k+\delta)T_c) + \eta((k+\delta)T_c) \qquad (4.17b)$$

where for $\delta \epsilon [0,1]$ Φ_δ, Γ_δ and $\omega_\delta(kT_c)$ are given by

$$\Phi_\delta = e^{A\delta} \qquad (4.17c)$$

$$\Gamma_\delta = \int_0^\delta e^{A(\delta-\sigma)}B\,d\sigma \qquad (4.17d)$$

$$\omega_\delta(kT_c) = \int_0^\delta e^{A(\delta-\sigma)}\omega(\sigma)\,d\sigma \qquad (4.17e)$$

It has been shown in lemma 2.3 of chapter 2 that if the sequence $\{u(kT_c)\}$ is a zero mean WSS sequence then u(t) defined in (4.1a) is zero-mean wide-sense cyclostationary (WSCS) and as a consequence the continuous vector state $\{x(t)\}$ and output $\{y(t)\}$ are also WSCS. It is a simple extension to show that $\{x(t)\}$ is also WSCS when $u(kT_c)$ is given by (4.7b) for a constant gain matrix G.

Define the covariance matrix $X(\delta,G)$ as

$$X(\delta,G) \triangleq \xi\{x((k+\delta)T_c)x'((k+\delta)T_c)\} \tag{4.18a}$$

From (4.17a), we obtain the covariance equation

$$X(\delta,G) = \bar{\Phi}_\delta X(0,G)\bar{\Phi}'_\delta + \Omega_\delta \tag{4.18b}$$

where $X(0,G)$ $(=X(1,G))$ is given by $\delta=1$ in (4.18a–b), and where Ω_δ is the covariance matrix of $\{\omega_\delta(kT_c)\}$ defined in (4.17e) and is given by

$$\Omega_\delta = \int_0^\delta e^{A(\delta-\sigma)}\Omega_c e^{A'(\delta-\sigma)}d\sigma \tag{4.18c}$$

The problem of discrete control has traditionally been treated by considering only $X(0,G)$ while ignoring the periodic properties of $X(\delta,G)$. These characteristics are the statistical equivalent of the deterministic phenomena of intersample ripple [Tou (1959)].

Consider the following quadratic cost function

$$J_f(\{\beta_i\},u) = \xi\{\sum_{k=-n}^{m} I(kT_c,\{\beta_i T_c\}) + x'((m+1)T_c)Q_0 x((m+1)T_c)\} \tag{4.19a}$$

where $u=\{u(kT_c)\}$ and where

$$I(kT_c,\{\beta_i T_c\}) = \sum_{i=1}^{N} [x'((k+\beta_i)T_c)Q_i x((k+\beta_i)T_c) + u'(kT_c)Ru(kT_c)] \tag{4.19b}$$

with $R>0$, $Q_i\geq 0$ and $\beta_i\in[0,1]$ for all $i=1,2,\ldots,N$. The terms with $x((k+\beta_i)T_c)$ in (4.19b) indicate the inclusion of intersample behavior in the design. A simplified form of (4.19b) can be written as

$$I(kT_c,\beta T_c) = x'((k+\beta)T_c)Q_1 x((k+\beta)T_c) + x'((k+1)T_c)Q_2 x((k+1)T_c) + u'(kT_c)Ru(kT_c)) \tag{4.19c}$$

where $\beta_1=\beta\in(0,1)$. We shall restrict ourselves to this form. Note that when $Q_1=0$, (4.19c) is in standard LQG form. The minimization of $J_f(\beta=\delta,u)$ can be reformulated as a standard LQG design, as later developed in lemma 4.3, resulting in the linear control law

$$u(kT_c) = -G_\delta(kT_c)x(kT_c) \tag{4.20}$$

The resulting cost at other intersample times $\sigma\epsilon[0,1]$ is given by $J_f(\sigma,u=G_\delta x)$. In practice, the worst intersample cost limits the control system performance. That is, for the fixed control law defined in (4.20) we are interested in

$$\max_{\sigma\epsilon[0,1]} \quad J_f(\sigma, u=G_\delta x) \tag{4.21a}$$

The overall minimax design philosophy is then to seek the optimal control law that achieves

$$\min_{\delta\epsilon[0,1]} \{ \max_{\sigma\epsilon[0,1]} \quad J_f(\sigma, u=G_\delta x)\} \tag{4.21b}$$

Definition 4.1 The discrete *minimax quadratic regulator* (MMQR) is defined by the gain $\{G_\delta(kT_c)\}$ in (4.20) such that the minimax criterion (4.21b) is accomplished. ·

□□□

As we have said, the problem of minimizing the cost $J_f(\beta=\delta,u)$ for a fixed δ is well defined. However $J_f(\sigma,u=G_\delta x)$ for fixed δ may possess more than one maximizing value $\bar{\sigma}$ in (4.19a). A (local) optimum solution of the MMQR criterion can be sought by using the following procedure.

Algorithm 4.1

0. Initialize j=0, $\delta_j=\delta_0\epsilon[0,1]$. (say $\delta_0=0.5$)
1. Obtain the optimal control gain G_{δ_j} defined in (4.20) (see lemma 4.1)
2. Find $\bar{\sigma}_j$ such that

$$f(\bar{\sigma}_j,\delta_j) \triangleq J_f(\bar{\sigma}_j, u=G_{\delta_j}x) = \max_{\sigma_j\epsilon[0,1]} J_f(\sigma_j, u=G_{\delta_j}x) \tag{4.22}$$

3. If j<2, set j=j+1, $\delta_j=\bar{\sigma}_{j-1}$ and go to step 1, else
4. If, for 'sufficiently small' ϵ_1 and ϵ_2,

   i. $|\delta_j - \delta_{j-1}| < \epsilon_1$, $|\bar{\sigma}_j - \bar{\sigma}_{j-1}| < \epsilon_1$ and

   ii. $|f(\bar{\sigma}_j,\delta_j) - f(\bar{\sigma}_{j-1},\delta_{j-1})| < \epsilon_2|f(\bar{\sigma}_{j-1},\delta_{j-1})|$

terminate the algorithm, otherwise set j=j+1, $\delta_j=\bar{\sigma}_{j-1}$ and go to step 1.

□□□

Note that in step 3 and step 4 of algorithm 4.1, the new choice of δ_j is $\bar{\sigma}_{j-1}$ (ie. the intersample time where the worst cost occurs). This approach seems justified since we do wish to minimize the worst (intersample) cost. Furthermore, exhaustive numerical evaluation in the examples shows that the optimum choice of δ does in fact coincide with the optimum $\bar{\sigma}$. Even though convergence of the algorithm cannot be justified analytically, the optimal solution in all examples was achieved in several iterations (<5). Values of $\varepsilon_1=\varepsilon_2=10^{-2}$ were used in all cases.

In order to find the solution to the MMQR problem we need to establish sufficient conditions under which the two sub-problems (ie. step 1 and step 2 of the algorithm) are well defined. We first examine the problem of finding the control sequence $\{u(kT_c)\}$ which minimizes $J_f(\delta,u)$.

The representation of the cost function (4.19a), (4.19c) and the plant (4.17a-e) can be transformed into a standard form by direct manipulations. The result is stated in the following lemma.

Lemma 4.3 Consider the equivalent discrete-time representations (4.4a-e) and (4.17a-e). Consider also the quadratic performance index $J_f(\beta=\delta,u)$ in (4.19a) with $I(kT_c,\beta T_c)$ defined in (4.19c). Then, the discrete equation (4.14a) for each $\delta\in[0,1]$ can be rewritten as

$$x((k+1)T_c) = \tilde{\Phi}_\delta x(kT_c) + \Gamma\tilde{u}(kT_c) + \omega(kT_c) \tag{4.23a}$$

where

$$\tilde{u}(kT_c) = u(kT_c) + R_\delta^{-1}M_\delta x(kT_c) \tag{4.23b}$$

$$\tilde{\Phi}_\delta = \Phi - \Gamma R_\delta^{-1}M_\delta \tag{4.23c}$$

$$R_\delta = R + \Gamma_\delta' Q_1 \Gamma_\delta + \Gamma' Q_2 \Gamma \tag{4.23d}$$

$$M_\delta = \Gamma_\delta' Q_1 \Phi_\delta + \Gamma' Q_2 \Phi \tag{4.23e}$$

where Φ, Γ and are defined by (4.4c-d) and $\{\omega(kT_c)\}$ is defined in (4.4a).

Furthermore, the quadratic cost $J_f(\delta,u)$ can be expressed as

$$J_f(\delta,u) = \xi\{\sum_{k=-n}^{m}[x'(kT_c)\tilde{Q}_\delta x(kT_c) + \tilde{u}'(kT_c)R_\delta\tilde{u}(kT_c)]\} + tr(Q_1\Omega_\delta+Q_2\Omega) \tag{4.24a}$$

where Ω is defined by (4.5a), Ω_δ is the covariance matrix of $\{\omega_\delta(kT_c)\}$ defined in (4.17e) and is given by (4.18c), and tr(.) denotes the trace of a matrix, and where

$$\tilde{Q}_\delta = \Phi_\delta' Q_1 \Phi_\delta + \Phi' Q_2 \Phi - M_\delta' R_\delta^{-1} M_\delta \qquad (4.24b)$$

Proof: First, substitution of $u(kT_c)$ in (4.17a) using (4.23b) results in (4.23a). Then in (4.19c), substitute $x((k+\delta)T_c)$ and $x((k+1)T_c)$ using (4.17a) and complete the squares to obtain

$$I(kT_c,\delta T_c) = x'(kT_c)\tilde{Q}_\delta x(kT_c) + \tilde{u}'(kT_c)R_\delta \tilde{u}(kT_c) + \theta(\omega_{1,2}(kT_c)) + \\ + \omega_\delta'(kT_c)Q_1\omega_\delta(kT_c) + \omega_1'(kT_c)Q_2\omega_1(k)$$

where $\tilde{Q}_\delta$ is given by (4.24b). $I(k,\delta)$ is then substituted into (4.19a) to give $J_f(\delta,u)$ in (4.24a). The term $\theta(\omega_{1,2}(kT_c))$ contains only cross product terms of the form $\omega_1(kT_c)x(kT_c)$, $\omega_1(kT_c)u(kT_c)$, $\omega_\delta(kT_c)x(kT_c)$ and $\omega_\delta(kT_c)u(kT_c)$. Consequently, since the process disturbance is assumed to be independent, $\xi\{\theta(\omega_{1,2}(kT_c))\}=0$. The term tr(.) in (4.24a) is obtained from the last two terms of the right hand side of $I(kT_c,\delta T_c)$ and using the fact that $\xi\{x'Px\}=\mathrm{tr}(P\xi\{xx'\})$.

□□□

Now in (4.23b), let

$$\tilde{u}(kT_c) = -\tilde{G}_\delta(kT_c)x(kT_c) \qquad (4.25a)$$

where

$$\tilde{G}_\delta(kT_c) = (R_\delta+\Gamma' S_\delta((k+1)T_c)\Gamma)^{-1}\Gamma' S_\delta((k+1)T_c)\tilde{\Phi}_\delta \qquad (4.25b)$$

and where $S_\delta(kT_c)$ (≥ 0) satisfies the Riccati difference equation

$$S_\delta(kT_c) = \tilde{\Phi}_\delta' S_\delta((k+1)T_c)\tilde{\Phi}_\delta + \tilde{Q}_\delta \\ - \tilde{\Phi}_\delta' S_\delta((k+1)T_c)\Gamma(R_\delta+\Gamma' S_\delta((k+1)T_c)\Gamma)^{-1}\Gamma' S_\delta((k+1)T_c)\tilde{\Phi}_\delta \qquad (4.25c)$$

Then from (4.23b), we get the control law (4.20) where

$$G_\delta(kT_c) = \tilde{G}_\delta(kT_c) + R_\delta^{-1}M_\delta \qquad (4.26)$$

For the *infinite horizon* problem (ie. $m,n\to\infty$) the control gain $\tilde{G}_\delta$ is time invariant and is given by the steady state solution of (4.25b) and (4.25c), provided such solution exists. Consequently, the control laws $u(kT_c)$ and $\tilde{u}(kT_c)$ defined in (4.20) and (4.25a), respectively become

$$u(kT_c) = -G_\delta x(kT_c) \tag{4.27a}$$

$$\tilde{u}(kT_c) = -\tilde{G}_\delta x(kT_c) \tag{4.27b}$$

where

$$G_\delta = \tilde{G}_\delta + R_\delta^{-1} M_\delta \tag{4.27c}$$

The resulting optimal control minimizes the limit of $J_f(\delta,u)$; that is

$$J(\delta,u) = \lim_{m\to\infty} \frac{1}{2m} \xi\{ \sum_{k=-m}^{m} I(k,\delta)\} \tag{4.28}$$

where $I(k,\delta)$ is defined by $\beta=\delta$ in (4.19c). Henceforth we restrict ourselves to the infinite horizon problem.

Now for each fixed δ, $u(kT_c)$ defined by (4.27a-c) minimizes $J(\delta,u)$ in (4.28) if and only if the steady state solution of the Riccati equation is *stabilizing* (ie. $\tilde{\Phi}_\delta - \Gamma\tilde{G}_\delta = \Phi - \Gamma G_\delta$ has all eigenvalues strictly inside the unit circle). A sufficient condition for the existence of a (unique) stabilizing solution is the controllability of $\{\tilde{\Phi}_\delta,\Gamma\}$ and the observability of $\{\tilde{\Phi}_\delta,\tilde{Q}_\delta^{\frac{1}{2}}\}$.

Given a minimal continuous time plant (4.1a-b), from lemma 2.1 of chapter 2 we know that the discrete realization (4.17a-e) for $\delta=1$ (or the discrete realization (4.4a-b)) is minimal if for any integer n, $\mathrm{Im}(\lambda_m(A)-\lambda_k(A))\neq 2\pi n$ whenever $\mathrm{Re}\lambda_m(A)=\mathrm{Re}\lambda_k(A)$ where $\mathrm{Im}(\rho)$ and $\mathrm{Re}(\rho)$ denote the imaginary and real part respectively of the complex number ρ. We assume the discrete model (4.4a-e) is minimal. (Else perturb the sampling period to $T_c=1+\epsilon$). Thus, we may assume $\{\Phi,\Gamma\}$ is controllable. The controllability of the pair $\{\tilde{\Phi}_\delta,\Gamma\}$ can be established by looking at $\tilde{\Phi}_\delta$ as being a state feedback realization of Φ and Γ, the result is presented in the following lemma.

Lemma 4.4 Consider the discrete realizations (4.4a-e) and (4.17a-e). Assume the pair $\{\Phi,\Gamma\}$ where Φ and Γ are given by (4.4c-d) is controllable. Then, the representation $\{\tilde{\Phi}_\delta,\Gamma\}$ where $\tilde{\Phi}_\delta$ is defined in (4.23c) is controllable for all $\delta\in(0,1]$.

Proof: Consider the discrete time description (4.4a); that is

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \tag{4.29a}$$

where $\{\Phi,\Gamma\}$ $(=\{\Phi_1,\Gamma_1\})$ is controllable. Assume the control law $u(kT_c)$ is governed by

$$u(kT_c) = -R_\delta^{-1}M_\delta x(kT_c) \tag{4.29b}$$

It is clear from (4.29a) and (4.29b) that $\{\tilde{\Phi}_\delta,\Gamma\}$ where $\tilde{\Phi}_\delta$ is defined in (4.23c) is a state feedback realization of $\{\Phi,\Gamma\}$. Hence $\{\tilde{\Phi}_\delta,\Gamma\}$ is also controllable [Brockett (1970)].

□□□

The following result establishes the observability of $\{\tilde{\Phi}_\delta,\tilde{Q}_\delta^{\frac{1}{2}}\}$ in the special case when $Q_2=0$ in (4.19c). We have been unable to establish a similar result in the general case.

Lemma 4.5 Consider the discrete representation (4.23a–e) with $Q_1\geq 0$, $Q_2=0$ and $R>0$. Suppose the pair $\{\Phi,Q_1^{\frac{1}{2}}\}$ is observable. Then, with $\tilde{Q}_\delta$ defined in (4.24b), the pair $\{\tilde{\Phi}_\delta,\tilde{Q}_\delta^{\frac{1}{2}}\}$ is observable for all $\delta\in[0,1]$.

Proof: Let

$$\underset{\sim}{G}_\delta = \Gamma R_\delta^{-1}\Gamma_\delta' Q_1^{\frac{1}{2}} \tag{4.30}$$

then from (4.23c), we obtain

$$\tilde{\Phi}_\delta = \Phi - \underset{\sim}{G}_\delta Q_1^{\frac{1}{2}}\Phi_\delta \tag{4.31}$$

Now if $\{\Phi,Q_1^{\frac{1}{2}}\}$ is observable so too is $\{\Phi,Q_1^{\frac{1}{2}}\Phi_\delta\}$. To see this, note that Φ_δ is nonsingular for all $\delta\in[0,1]$ and Φ and Φ_δ commute. The observability of $\{\tilde{\Phi}_\delta,Q_1^{\frac{1}{2}}\Phi_\delta\}$ then follows from (4.31). Now after substitution of R_δ and M_δ using (4.23d–e), $\tilde{Q}_\delta$ in (4.24b) can be represented as

$$\tilde{Q}_\delta = \Phi_\delta' Q_1^{\frac{1}{2}}\{I-Q_1^{\frac{1}{2}}\Gamma_\delta(R+\Gamma_\delta' Q_1^{\frac{1}{2}}Q_1^{\frac{1}{2}}\Gamma_\delta)^{-1}\Gamma_\delta' Q_1^{\frac{1}{2}}\}Q_1^{\frac{1}{2}}\Phi_\delta \tag{4.32}$$

An application of the matrix inversion lemma [Kailath (1980)] to the term $\{.\}$ in (4.32) gives

$$\tilde{Q}_\delta = \Phi_\delta' Q_1^{\frac{1}{2}}(I+Q_1^{\frac{1}{2}}\Gamma_\delta R^{-1}\Gamma_\delta' Q_1^{\frac{1}{2}})^{-1}Q_1^{\frac{1}{2}}\Phi_\delta \tag{4.33}$$

Now write

$$\Psi_\delta = (I+Q_1^{\frac{1}{2}}\Gamma_\delta R^{-1}\Gamma_\delta' Q_1^{\frac{1}{2}})^{-1} \tag{4.34}$$

Then $R>0$ implies $\Psi_\delta>0$, and from (4.33) and (4.34), we obtain

$$\tilde{Q}_\delta^{\frac{1}{2}} = \Psi_\delta^{\frac{1}{2}}Q_1^{\frac{1}{2}}\Phi_\delta \tag{4.35}$$

Now the observability matrix of the pair $\{\tilde{F}_\delta, \tilde{Q}_\delta^{\frac{1}{2}}\}$ is defined by

$$O_\delta = \begin{bmatrix} \tilde{Q}_\delta^{\frac{1}{2}} \\ \tilde{Q}_\delta^{\frac{1}{2}}\tilde{F}_\delta \\ \vdots \\ \tilde{Q}_\delta^{\frac{1}{2}}\tilde{F}_\delta^{N-1} \end{bmatrix}$$

Substitution of $\tilde{Q}_\delta^{\frac{1}{2}}$ in O_δ using (4.35) implies

$$O_\delta = \Psi_\delta^{\frac{1}{2}} \underset{\sim}{O}_\delta$$

where $\underset{\sim}{O}_\delta$ represents the observability matrix for the pair $\{\tilde{\Phi}_\delta, Q_1^{\frac{1}{2}}\Phi_\delta\}$. Since $\Psi_\delta^{\frac{1}{2}}$ and $\underset{\sim}{O}_\delta$ are full rank matrices, O_δ is also full rank.

□□□

Lemmas 4.3, 4.4 and 4.5 facilitate the following result which is useful for solving the minimization of $J(\delta,u)$ in (4.28).

Theorem 4.1 Consider the discrete-time representations (4.4a-e) and (4.17a-e) of the minimal continuous-time system (4.1a-b) under the action of the pulse amplitude modulated control (4.2b-c) with a sampling period T_c. Without loss of generality assume the representation (4.4a-e) is minimal. For any $\delta \in [0,1]$ consider the performance index $J(\delta,u)$ defined by (4.28) where $I(kT_c, \beta T_c = \delta T_c)$ is defined by (4.19c) with $Q_1 \geq 0$, $Q_2 = 0$ and $R > 0$. Assume the pair $\{\Phi, Q_1^{\frac{1}{2}}\}$ is observable.

Then, $\{\tilde{\Phi}_\delta, \tilde{Q}_\delta^{\frac{1}{2}}\}$ where $\tilde{\Phi}_\delta$ and $\tilde{Q}_\delta$ are given respectively by (4.23c) and (4.24b) is observable. Furthermore, the optimal control sequence $\{u(kT_c)\}$ which minimizes $J(\delta,u)$ is defined by (4.27a-c) where R_δ and M_δ are given by (4.23d-e) and where the stabilizing $\tilde{G}_\delta$ is given by the steady state solution of (4.25b-c).

□□□

Theorem 4.1 together with lemmas 4.3, 4.4 and 4.5 provides sufficient conditions under which the sub-problem in step 1 of the algorithm 4.1 is well defined. We now examine the solution of the sub-problem in step 2 of the algorithm 4.1.

The maximization problem in (4.21a) which is restated in step 2 of

algorithm 4.1 has no analytical solution and so we adopt a numerical approach. Any line search method such as quadratic fit or golden section search (see for example [Polak (1971) or Gill et. al (1981)]) is suitable for attacking this problem. The following result is needed in this context.

Lemma 4.6 Consider the discrete-time plant (4.17a-e) and consider as well the quadratic performance index $J(\delta,u)$ in (4.28) where $I(k,\beta=\delta)$ is defined by (4.19c) with $Q_1 \geq 0$, $Q_2=0$ and $R>0$. Consider also the optimal control law $u(kT_c)$ in (4.27a-c) which is derived by theorem 4.1 for a fixed $\delta \in [0,1]$.

Then, for any $\sigma \in [0,1]$ the intersample cost is given by

$$J(\sigma, u=G_\delta x) = tr[X(\sigma,G_\delta)Q_1 + G_\delta' R G_\delta X(1,G_\delta)] \qquad (4.36a)$$

where the gain G_δ is given by (4.27c) and where the (unique) symmetric $X(\sigma,G_\delta)$ satisfies the discrete Lyapunov equation

$$X(\sigma,G_\delta) = (\Phi_\sigma - \Gamma_\sigma G_\delta)X(1,G_\delta)(\Phi_\sigma - \Gamma_\sigma G_\delta)' + \Omega_\sigma \qquad (4.36b)$$

where Φ_δ, Γ_δ and Ω_δ are respectively given by (4.17c-d) and (4.18c).

Proof: Substitution of the optimal control $u(kT_c)$ in (4.17a) using (4.27b-c) gives the closed-loop representation for any intersample time $\sigma \in [0,1]$

$$x((k+\sigma)T_c) = (\Phi_\sigma - \Gamma_\sigma G_\delta)x(kT_c) + \omega_\sigma(kT_c) \qquad (4.37)$$

Now define

$$X(\sigma,K_\delta) \triangleq \xi\{x(k+\sigma)x'(k+\sigma)\}$$

Then, the covariance equation (4.36b) follows form (4.37). Substitution of $u(kT_c)$ in $I(kT_c,\beta T_c=\sigma T_c)$ defined by (4.19c) using (4.27b) gives

$$I(kT_c,\sigma T_c) = x'(k+\sigma)Q_1 x(k+\sigma) + (k+\sigma)G_\delta' R G_\delta x(kT_c) \qquad (4.38a)$$

From (4.28), the cost at any intersample time $\sigma \in [0,1]$ is given by

$$J(\sigma, u=G_\delta x) = \lim_{m\to\infty} \frac{1}{2m} \xi\{\sum_{k=-m}^{m} I(k,\sigma)\} \qquad (4.38b)$$

where $I(k,\sigma)$ is given in (4.38a). Interchange the expectation operator with the

summation sign in (4.38b) and use the fact that $\xi\{x'Px\}=tr[X(1,K_\delta)P]$ to obtain (4.36a).

□□□

A termination condition is necessary for the line search method. One possibility is to stop the iteration when the search interval is less than ϵ for some $\epsilon>0$. However, the choice of ϵ depends on the flatness of $J(\sigma,u=G_\delta x)$. In the examples, we tested for a (local) maximum at $\sigma=\bar{\sigma}$ by numerical evaluation of the derivative with respect to σ of $J(\sigma,u=G_\delta x)$ using the formulae defined in the following lemma.

Lemma 4.7 Consider the discrete-time model (4.17a-e) and the intersample cost $J(\sigma,u=G_\delta x)$ in (4.36a). Then, the derivative of the index $J(\sigma,u=G_\delta x)$ with respect to σ is given by

$$\frac{\partial J(\sigma,u=G_\delta x)}{\partial \sigma} = tr\left[\frac{\partial X(\sigma,G_\delta)}{\partial \sigma} Q_1\right] \tag{4.39a}$$

where

$$\frac{\partial X(\sigma,G_\delta)}{\partial \sigma} = (\Phi_\sigma - \Gamma_\sigma G_\delta)X(1,G_\delta)(A - BG_\delta)'\Phi_\sigma'$$

$$+ \Phi_\sigma(A - BG_\delta)X(1,G_\delta)(\Phi_\sigma - \Gamma_\sigma G_\delta)' + A\Omega_\sigma + \Omega_\sigma A' + \Omega_c \tag{4.39b}$$

where the matrices A and B are the continuous system matrices defined in (4.1a-b) and the matrix Ω_c is the covariance of the continuous process $\omega(t)$ defined in (4.2a).

Proof: On the right hand side of (4.36a) only the first term is dependent on σ. Therefore, the derivative of the index $J(\sigma,G_\delta)$ with respect to σ can easily be shown to be given by (4.39a). From (4.17c-d) and (4.18c) the following relations which are useful for obtaining the derivative of the covariance function $X(\sigma,G_\delta)$ in (4.36b) can be derived

$$\frac{\partial \Phi_\sigma}{\partial \sigma} = A\Phi_\sigma \tag{4.40a}$$

$$\frac{\partial \Gamma_\sigma}{\partial \sigma} = B\Phi_\sigma \tag{4.40b}$$

$$\frac{\partial \Omega_\sigma}{\partial \sigma} = \Omega_c + A\Omega_\sigma + \Omega_\sigma A' \qquad (4.40c)$$

Taking the derivative of the covariance matrix $X(\sigma, G_\delta)$ and make use of the relations (4.40a–c) gives (4.39b).

□□□

4.4 MINIMAX OUTPUT VARIANCE REGULATION

In this section, we consider the special case of output variance regulation of a single–input single–output (SISO) plant (4.1a–b). That is, we seek to minimize the criterion (4.28) with

$$I(kT_c, \beta T_c) = ay^2((k+1)T_c) + by^2((k+\beta)T_c) + Ru^2(kT_c) \qquad (4.41)$$

The classical *minimum variance regulator* (MVR) [Åström (1970)] minimizes (4.20) and (4.41) for a=1, b=0 and R=0. The value of R=0 in the MVR cost function means there is no weighting factor applied on the manipulated variable $u(kT_c)$ when the minimization is carried out. In many cases, this results in very large input energy for regulation. The large deviations in the sequence $\{u(kT_c)\}$ may produce large variations in the periodic output variance in the intersample period as demonstrated in [de Souza and Goodwin (1984)].

Several methods have been suggested to overcome the large variations in $\{u(kT_c)\}$. The so called *cheap* (infinite horizon) *control* is actually an addition of a small weighting factor R on the squared input in the classical MVR criterion. This criterion can be written in the form of (4.28) with

$$I(kT_c, \beta=0) = y^2((k+1)T_c) + Ru^2(kT_c)$$

which corresponds to a=1, b=0 and R 'sufficiently small' in (4.41). It has been shown that with the right choice of a weighting factor R, the large variation of $\{u(kT_c)\}$ can be reduced to an acceptable level. Nevertheless, the choice of the weighting factor R is somewhat artificial. We show this by means of some illustrative examples.

An alternative solution for R=0 as suggested in [Toivonen (1983)] places a magnitude constraint on the control signal of the form

$$u(kT_c) = \text{sat}(-Kx;\zeta,\Theta)$$

where K is the feedback gain and the function sat(.;.,.) is defined by

$$\text{sat}(z;\zeta,\Theta) = \begin{cases} \zeta & \text{if } z<\zeta \\ z & \text{if } \zeta<z<\Theta \\ \Theta & \text{if } z>\Theta \end{cases}$$

This approach is justifiable on practical ground but it leads to a difficult if not intractible nonlinear dynamic programming solution.

Another possible solution is to minimize the variance of the plant output subject to a bound ν on the input variance; that is

$$\xi\{u^2(kT_c)\} \leq \nu$$

This alternative is discussed in [Mäkilä et. al (1984)]. In this approach, an appropriate value of $\nu>0$ must first be chosen to guarantee the existence of a stabilizing solution.

Another method which takes into account the behavior of the system in intersample time by considering an integral criterion rather than a summation as in (4.28) is discussed in [Kwakernaak and Sivan (1972)]. Specifically, consider the minimization of the continuous function J_{ksc} defined by

$$J_{ksc} = \lim_{n\to\infty} \frac{1}{2n} \xi\{ \int_{-n}^{n} y^2(t)dt\} \tag{4.42}$$

Then, for the pulse amplitude modulated control (4.2b–c), minimizing J_{ksc} subject to (4.1a–b) is equivalent to minimizing J_{ksd} subject to (4.4a–b) where J_{ksd} is given by; for a sampling period T_c

$$J_{ksd} = \lim_{m\to\infty} \frac{1}{2m} \xi\{ \sum_{k=-m}^{m} [x'(kT_c)Q_{ks}x(kT_c) + 2u'(kT_c)M_{ks}x(kT_c) + u'(kT_c)R_{ks}u(kT_c)]\} \tag{4.43a}$$

and where

$$Q_{ks} = \int_0^{T_c} \Phi'(\tau)L'L\Phi(\tau)d\tau \tag{4.43b}$$

$$M_{ks} = \int_0^{T_c} \Phi'(\tau)L'L\Gamma(\tau)d\tau \tag{4.43c}$$

$$R_{ks} = \int_0^{T_c} \Gamma'(\tau)L'L\Gamma(\tau)d\tau \tag{4.43d}$$

where $\Phi(\tau)$ and $\Gamma(\tau)$ are given by

$$\Phi(\tau) = e^{A\tau} \tag{4.43e}$$

$$\Gamma(\tau) = \int_0^{\tau} e^{A(\tau-\sigma)}Bd\sigma \tag{4.43f}$$

After completing square, the term inside the bracket [.] in (4.43a) becomes

$$x'(kT_c)\bar{Q}_{ks}x(kT_c) + \tilde{u}'(kT_c)R_{ks}\tilde{u}(kT_c) \tag{4.44a}$$

where

$$\bar{Q}_{ks} = Q_{ks} - M_{ks}R_{ks}^{-1}M_{ks}' \tag{4.44b}$$

$$\tilde{u}(kT_c) = u(kT_c) + R_{ks}^{-1}M_{ks}'x(kT_c) \tag{4.44c}$$

For simplicity we call this the KS (Kwakernaak and Sivan) approach.

Using the idea of the MMQR discussed in the previous section, we propose an alternative approach which is aimed at minimizing the maximum (or the worst) intersample variance of the output; that is

$$\min_{\delta\in[0,1]} \{ \max_{\sigma\in[0,1]} \xi\{y^2(k+\sigma)\}\} \tag{4.45}$$

where $\xi\{y^2(k+\sigma)\}$ is the variance of the plant output evaluated at the intersample time $\sigma\in[0,1]$ with control law defined in (4.19a-c) for fixed $\delta\in[0,1]$. Since the objective of this approach is to minimize the worst intersample variance, we then call this the *minimax* output *variance regulator* (MMVR). The intersample variance $\xi\{y^2(k+\sigma)\}$ can be represented by the criterion (4.20) with $I(k,\beta=\sigma)$ defined by (4.41) for a=0, b=1 and R=0. As in lemma 3.1, the MMVR problem can be transformed into a standard LQG case. The result is stated as follows.

Lemma 4.8 Consider a minimal continuous-time SISO plant (4.1a-b) regulated by u(t) as defined in (4.2b-c) having a minimal equivalent discrete-time representation (4.4a-b). Then, criterion (4.21b) with $I(k,\beta)$ defined in (4.41) for a=0, b=1 and R=0 corresponds to $Q_1=L'L$, $Q_2=0$ and R=0 in Lemma 4.3. The weighting factors $\tilde{Q}_\delta$, M_δ and R_δ defined respectively in (4.24b) and (4.23d-e) become

$$\tilde{Q}_\delta = 0 \tag{4.46a}$$

$$R_\delta = (L\Gamma_\delta)^2 \tag{4.46b}$$

$$M_\delta = (L\Gamma_\delta)L\Phi_\delta \tag{4.46c}$$

Furthermore $R_\delta>0$ for all $\delta\epsilon[0,1]$.

Proof: The weighting factors R_δ and M_δ in (4.46b-c) were derived by direct substitution of Q_1, Q_2 and R in (4.23d-e). Substitution of R_δ and M_δ in (4.24b) using (4.46b-c) yields (4.46a).

□□□

Note that $R_\delta>0$ follows since it is assumed that there is no pure time delay in the continuous plant; that is $\Gamma_\delta\neq0$ for $\delta(0,1]$. From Lemma 4.3, for $\tilde{Q}_\delta=0$, it follows that provided $|\lambda_j(\tilde{\Phi}_\delta)|<1$ for all $1\leq j\leq n_x$ the optimal solution is achieved by setting $\tilde{u}(kT_c)=0$; that is

$$u(kT_c) = -R_\delta^{-1}M_\delta x(kT_c) \tag{4.47}$$

is optimal. However, if the stability requirement is *not* fulfilled, then the steady state solution of the Riccati equation has to be solved for $\tilde{Q}_\delta=0$. Unfortunately, in this case we do not have the observability of $(\tilde{\Phi}_\delta,\tilde{Q}_\delta^{\frac{1}{2}})$ as required in theorem 4.1. This problem motivates a modification of the performance index in the case of minimax output regulation. From the cost (4.41), it can be seen that the modification can be done in several ways. One possibility is to extend the idea of cheap control by considering a=0, b=1, $\beta=\delta$ and a 'small' R in the cost (4.41). In this case, direct manipulations of the weighting factors $\tilde{Q}_\delta$, M_δ and R_δ defined respectively in (4.24b) and (4.23d-e) yield

$$\tilde{Q}_\delta = \Phi_\delta'L'L\Phi_\delta - M_\delta'R_\delta^{-1}M_\delta \tag{4.48a}$$

$$M_\delta = (L\Gamma_\delta)L\Phi_\delta \tag{4.48b}$$

$$R_\delta = R + (L\Gamma_\delta)^2 \tag{4.48c}$$

Another alternative is to let R=0 and introduce a 'small' a>0 and $\beta=\delta$ in (4.41), the weighting factors $\tilde{Q}_\delta$, M_δ and R_δ defined respectively in (4.24b) and (4.23d-e) become

$$\tilde{Q}_\delta = a\Phi_1' L' L\Phi_1 + b\Phi_\delta' L' L\Phi_\delta - M_\delta' R_\delta^{-1} M_\delta \tag{4.49a}$$

$$M_\delta = a(L\Gamma_1)L\Phi_1 + b(L\Gamma_\delta)L\Phi_\delta \tag{4.49b}$$

$$R_\delta = a(L\Gamma_1)^2 + b(L\Gamma_\delta)^2 \tag{4.49c}$$

Notice for the infinite horizon problem that the instant $(k+1)T_c$ can be replaced by kT_c in (4.41) since in steady state the output variance is constant for all instants. Using this alternative design the weighting factors $\tilde{Q}_\delta$, M_δ and R_δ defined in (4.24b) and (4.23d-e) become

$$\tilde{Q}_\delta = aL'L \tag{4.50a}$$

$$M_\delta = b(L\Gamma_\delta)L\Phi_\delta \tag{4.50b}$$

$$R_\delta = b(L\Gamma_\delta)^2 \tag{4.50c}$$

Both (4.49a-c) and (4.50a-c) result in the same steady state control law.

Theorem 4.2 Consider a minimal SISO time-invariant system (4.1a-b) and the control law (4.2b-c) for a sampling period T_c with minimal discrete realizations (4.4a-e) and (4.9a-e). Then, for any $\delta \in [0,1]$, the infinite horizon performance index $J(\delta,u)$ in (4.28) where $I(k,\beta=\delta)$ is given by (4.41) with $\tilde{Q}_\delta$, M_δ and R_δ are given either by (4.48a-c), (4.49a-c) or (4.50a-c) is minimized by the control sequence $\{u(kT_c)\}$ defined in (4.27a-c) where the gain $\tilde{G}_\delta$ is given by the steady-state solution of (4.25b-c). Moreover, the closed loop system is asymptotically stable provided the pair $(\tilde{\Phi}_\delta, \tilde{Q}_\delta^{\frac{1}{2}})$ with $\tilde{\Phi}_\delta$ given by (4.23c) has no unobservable modes on the unit circle.

□□□

Notice it follows from Lemma 4.4 that $\{\tilde{\Phi}_\delta, \tilde{Q}_\delta^{\frac{1}{2}}\}$ is observable under the assumption that $\{\Phi, L=Q_1^{\frac{1}{2}}\}$ is observable when $\tilde{Q}_\delta$ is given by (4.48a). However, the pair $\{\tilde{\Phi}_\delta, \tilde{Q}_\delta^{\frac{1}{2}}\}$ where $\tilde{Q}_\delta$ is given either by (4.49a) or (4.50a) may be unobservable for some values of δ. For example, if $\tilde{Q}_\delta$ is given by (4.49a), the pair $\{\tilde{\Phi}_\delta, \tilde{Q}_\delta^{\frac{1}{2}}\}$ is

unobservable for both $\delta=0$ and $\delta=1$. If $\tilde{Q}_\delta$ is given by (4.50a), the pair $\{\tilde{\Phi}_\delta,\tilde{Q}_\delta^{\frac{1}{2}}\}$ is unobservable for $\delta=1$. The result concerning the existence of the (unique) stabilizing gain $\tilde{G}_\delta$ as discussed in the previous section is due to [Chan et. al (1984)].

We now illustrate the application of the MMVR design by means of some examples.

Example 4.1 Consider a scalar model (4.1a) with $A=-\alpha$, $B=C=1$. Assume, for simplicity no measurement noise (ie. $\Lambda=0$) and $\Omega_c=1$ in (4.2a). The discrete version of the model (4.1a) for a sampling period T_c is governed by (see (4.4a-e) and (4.5a-b))

$$x((k+1)T_c) = e^{-\alpha}x(kT_c) + \alpha^{-1}(1-e^{-\alpha})u(kT_c) + \omega_1(kT_c)$$

For any $\delta\epsilon[0,1]$, in theorem 4.1 with $Q_1=q\geq 0$, $Q_2=0$ and $R=r>0$, it is not difficult to show that $\tilde{u}(kT_c)=0$ is stabilizing for all $\alpha>0$. The periodic variance $X(\sigma,G_\delta)$ is given by

$$X(\sigma,G_\delta) = (e^{-\alpha\sigma}-\alpha^{-1}(1-e^{-\alpha\sigma})G_\delta)^2X(1,G_\delta) + \Omega_\sigma$$

where G_δ is the stabilizing gain for fixed δ. Analytically, it can be shown that; for all $\delta\epsilon[0,1]$

$$X(1,G_\delta) \geq X(\sigma,G_\delta)$$

In other words, the maximum is situated at the boundary ($\sigma=1$). From the algorithm 4.1 discussed in section 4.3 and definition 4.1, the MMQR criterion is given by (4.28) for $\delta=1$ which is in fact a standard LQG cost function.

In the case of output variance regulation, the MMVR criterion is given by (4.28) where $I(k,\beta=\delta)$ is given by (4.41) for $\beta=\delta=1$, $a=0$, $b=1$ and $R=0$. Hence, the MMVR and the MVR are identical. The implementation of the KS method results in a better average intersample output variance, but it gives a higher maximum output variance. For example for $A=\alpha=1$ and $T_c=1$sec, the KS gives the variance $\xi\{y^2(kT_c)\}=0.462$ whereas the MMVR gives $\xi\{y^2(kT_c)\}=\Omega_1=0.4323$. Fig.4.2 shows the intersample output variance for both designs.

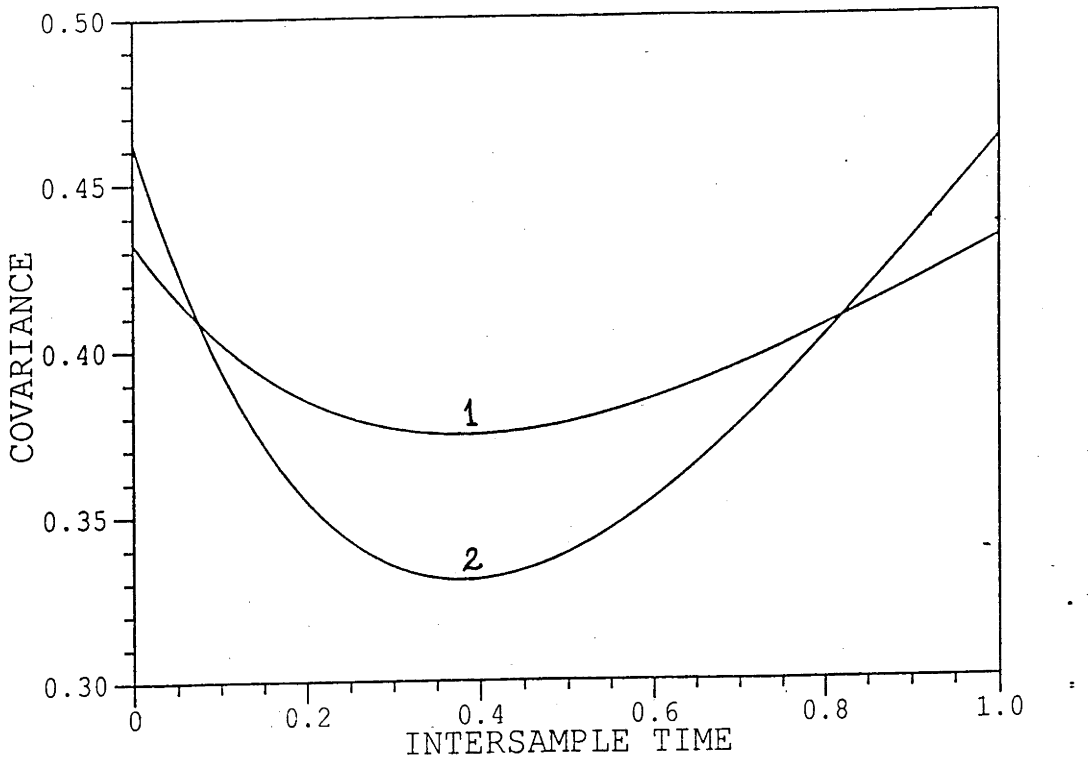


Fig.4.2 The intersample output variance of 1st order system corresponding to
1. MVR (=MMVR)
2. KS

Example 4.2 Consider a SISO second order plant (4.1a-b) where

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -\alpha \end{bmatrix} ; \; B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} ; \; L' = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} ; \; \Omega = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and where α=0.1, c_1=1 and c_2=0. This example is taken from [de Souza and Goodwin, (1984)]. With a sampling period T_c=1sec, the MVR gain was derived by minimizing the cost (4.28) where $I(k,\beta)$ is defined by (4.41) for a=1, b=0 and R=0. The resulting gain is G=[2.0697 1.9671]. Note that in this case $\tilde{u}(kT_c)$=0 is in fact a stabilizing control. The intersample output variance corresponding to the MVR gain was computed and is illustrated by the curve 1 in Fig.4.3. The corresponding control variance is $\xi\{u^2(kT_c)\}$=235.

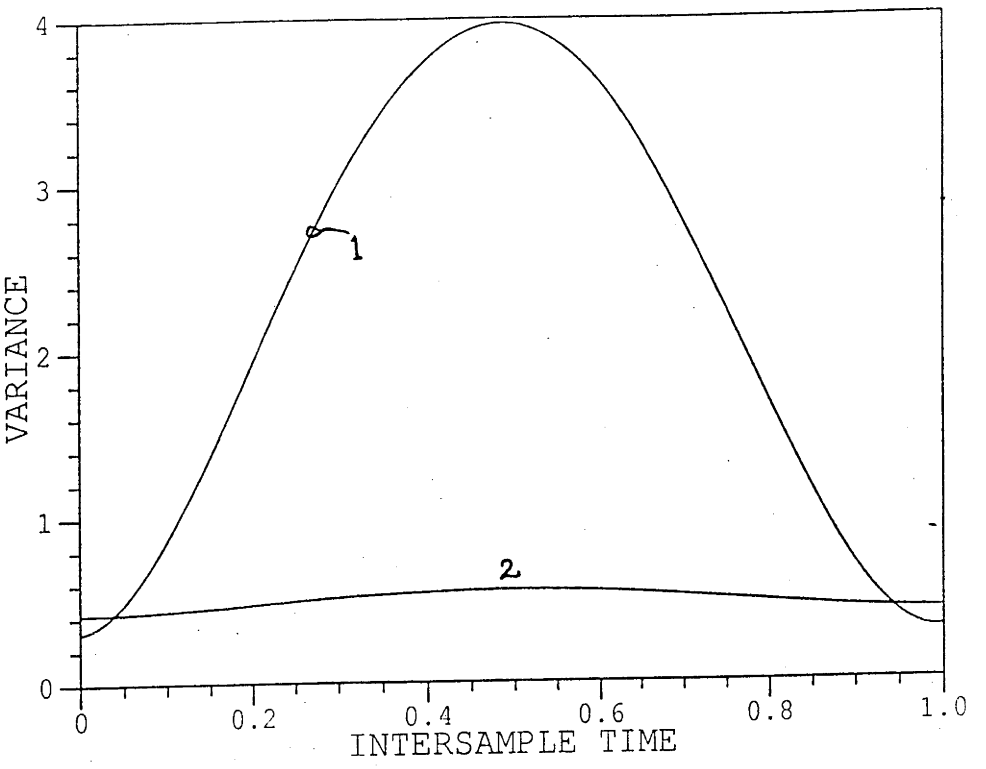


Fig.4.3 The intersample output variance of a 2nd order system corresponding to
1. MVR
2. MMVR for b/a=10^3

The cheap variance control corresponds to a=1, b=0 and R≠0 in (4.41). A weighting factor R=0.015 results in the cheap control gain G=[1.4502 1.6339]. The resulting intersample output variance is depicted by the curve 2 in Fig.4.4. The corresponding control variance is $\xi\{u^2(kT_c)\}$=8.6. In order to show the influence of the weighting factor R of the cheap control design on the closed-loop performance, the maximum variance of output ($=\sigma_{cm}(R)$) was computed for different values of R. The resulting $\sigma_{cm}(R)$ was then normalized by $\sigma_{cm}(R=10^{-6})$; that is

$$\bar{\sigma}_{cm}(R) = \frac{\sigma_{cm}(R)}{\sigma_{cm}(R=10^{-6})}$$

In Fig.4.5, we show the variations of the normalized maximum variance of output ($=\bar{\sigma}_{cm}(R)$) as R increases. The optimal weighting factor R of the cheap control design is given by $R^*\approx 0.015$ (ie. $\bar{\sigma}_{cm}(R^*=0.015)<\bar{\sigma}_{cm}(R)$ for the curve 1 in Fig.4.5).

To show the influence of the weighting factor R of the MMVR design defined by (4.49a-c) on the closed-loop performance, the maximum variance of output ($=\sigma_{mm}(r)$) was computed for different values of R. The resulting $\sigma_{mm}(R)$ was then normalized by $\sigma_{mm}(R=100)$; that is

$$\bar{\sigma}_{mm}(R) = \frac{\sigma_{mm}(R)}{\sigma_{mm}(R=100)}$$

The variations of the normalized maximum variance of output ($=\bar{\sigma}_{mm}(R)$) as R increases is illustrated in Fig.4.6. The optimal weighting factor R of the MMVR design is given by $R^*<0.1$ (ie. $\bar{\sigma}_{mm}(R^*<0.1)\leq\bar{\sigma}_{mm}(R)$ for the curve 1 in Fig.4.6).

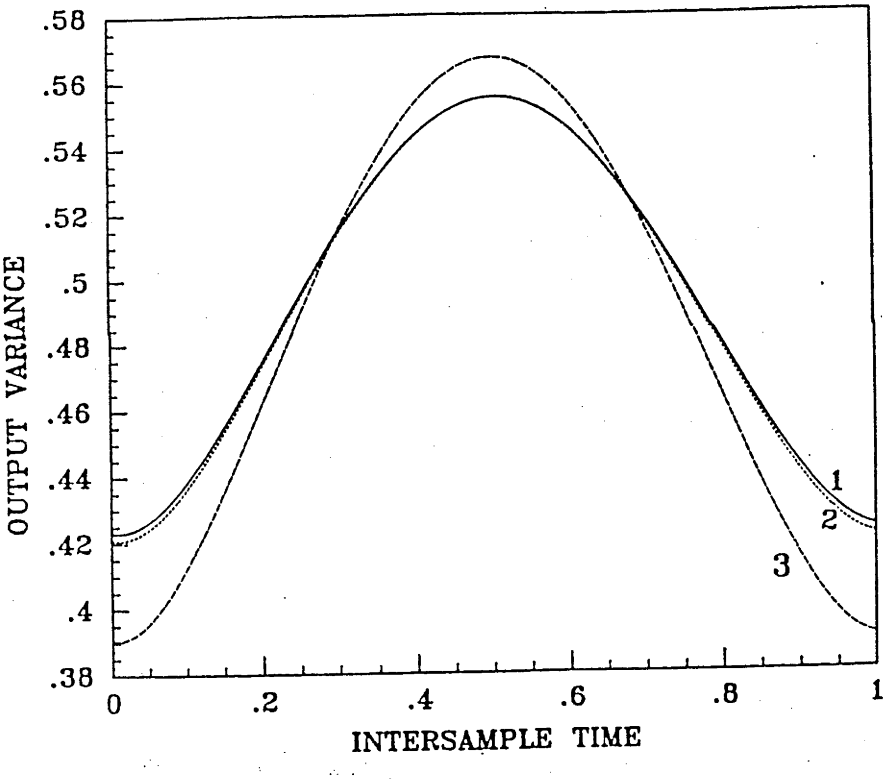


Fig.4.4 The intersample output variance of a 2nd order system corresponding to
1. MMVR for $b/a=10^3$
2. Cheap control for R=0.015
3. KS

Using the KS method described in (4.34)-(4.36), we obtain the KS gain G=[1.5679 1.7022]. The curve 3 of 4 Fig.4.4 shows the intersample output variance produced by this gain. The corresponding control variance is $\xi\{u^2(kT_c)\}=11.2$.

With δ_0=0.5 as an initial choice of δ, the step 2 of the algorithm produced the worst intersample cost which were situated around $\bar{\sigma}$=0.53 in all iterations. For fixed $\delta=\bar{\sigma}$=0.53 the control law u(kT_c) defined in (4.47) is not stabilizing. The value b/a=10^3 in (4.49a-c) was chosen to ensure the existence of the stabilizing solution of the step 1 of the algorithm. The algorithm converged to the (local) solution in 3 iterations. The resulted gain is G_δ=[1.4425 1.6291]. The variance of output in the intersample period is shown by the curve 1 in Fig.4.4 (or the curve 2 in Fig.4.3). For this example, we found that the intersample cost $J(\sigma,u=G_\delta x)$ is found to be within ±0.01% if its optimum value for $\delta\epsilon[0.42,0.57]$ and $\sigma\epsilon[0.51,0.56]$.

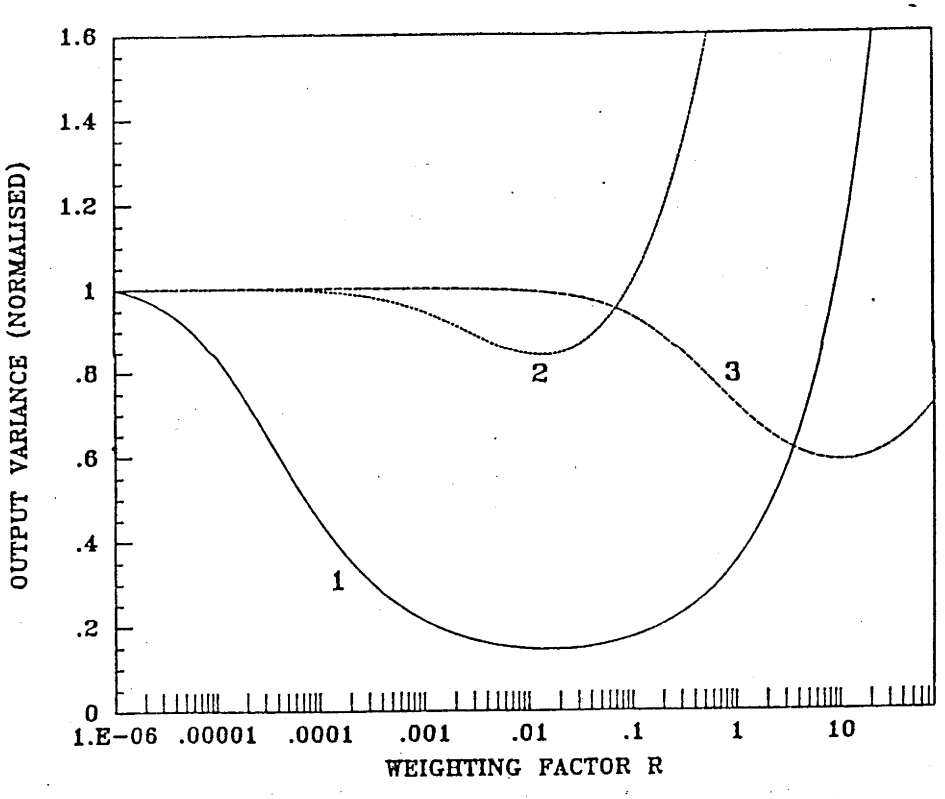


Fig.4.5 The maximum output variance versus control weighting R for cheap control in
1. Example 3.2
2. Example 3.3
3. Example 3.4

To show the influence of the weighting factors a and b of the MMVR design defined by (4.49a-c) (or (4.50a-c)), the maximum variance of the output (=σ_{mm}(b/a)) was computed for different ratios of b/a.

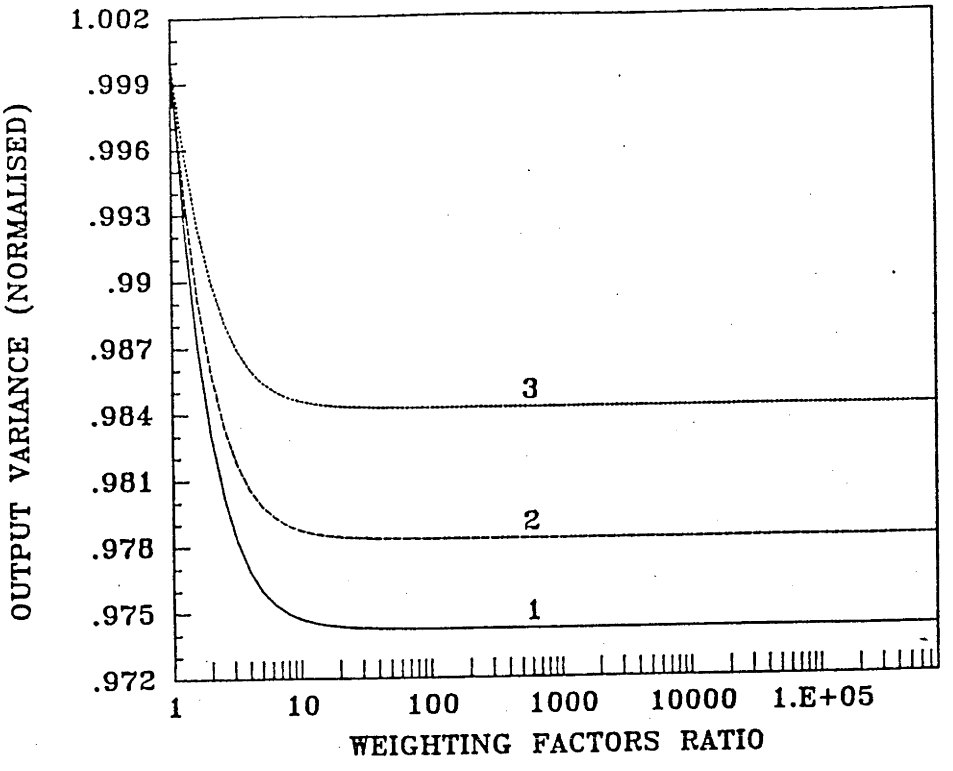


Fig.4.6 Maximum output variance versus control weighting R for MMVR in
1. Example 3.2
2. Example 3.3
3. Example 3.4

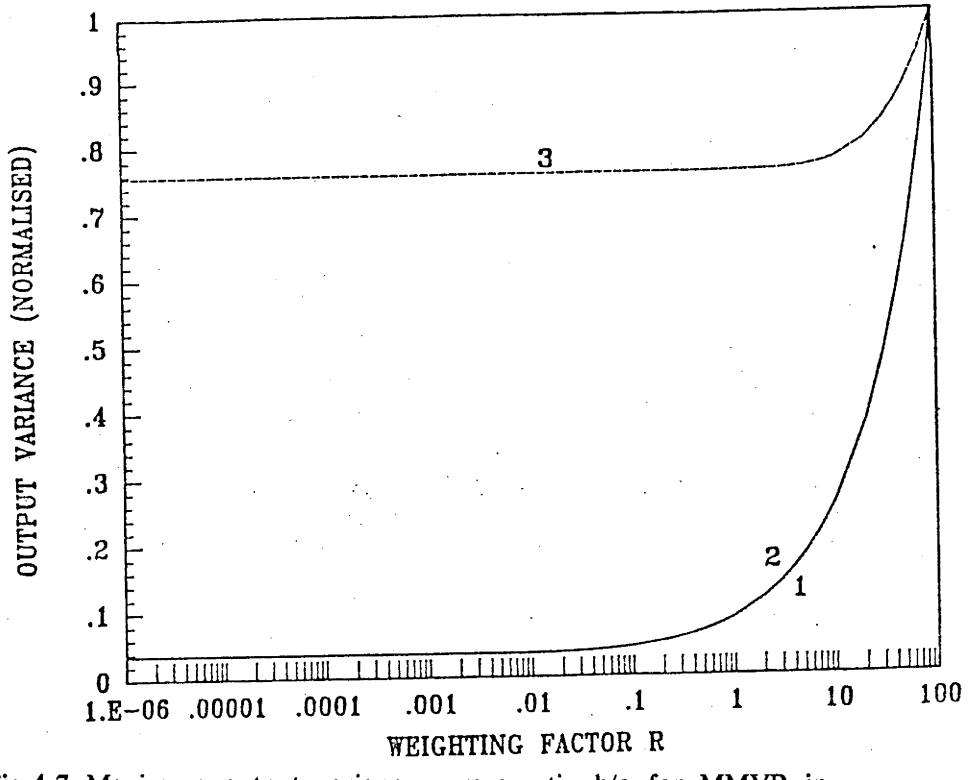


Fig.4.7 Maximum output variance versus ratio b/a for MMVR in

1. Example 3.2
2. Example 3.3
3. Example 3.4

The resulting $\sigma_{mm}(b/a)$ was then normalized by $\sigma_{mm}(b/a=1)$; that is

$$\bar{\sigma}_{mm}(b/a) = \frac{\sigma_{mm}(b/a)}{\sigma_{mm}(b/a=1)}$$

The variance of the normalized maximum variance of output ($=\bar{\sigma}_{mm}(b/a)$) as b/a increases is illustrated in Fig.4.7. The optimal ratio b/a of the MMVR design is given by $b/a^*>0$ (ie. $\bar{\sigma}_{mm}(b/a^*>10)<\sigma_{mm}(b/a)$ for the curve 1 in Fig.4.7).

Example 4.3 Consider again the SISO model used in example 4.2 where now $\alpha=0$. The output of the model is considered for three different choices of output vector C and for each given value of C the MVR gain were computed for a sampling period T_c=1sec; the result is

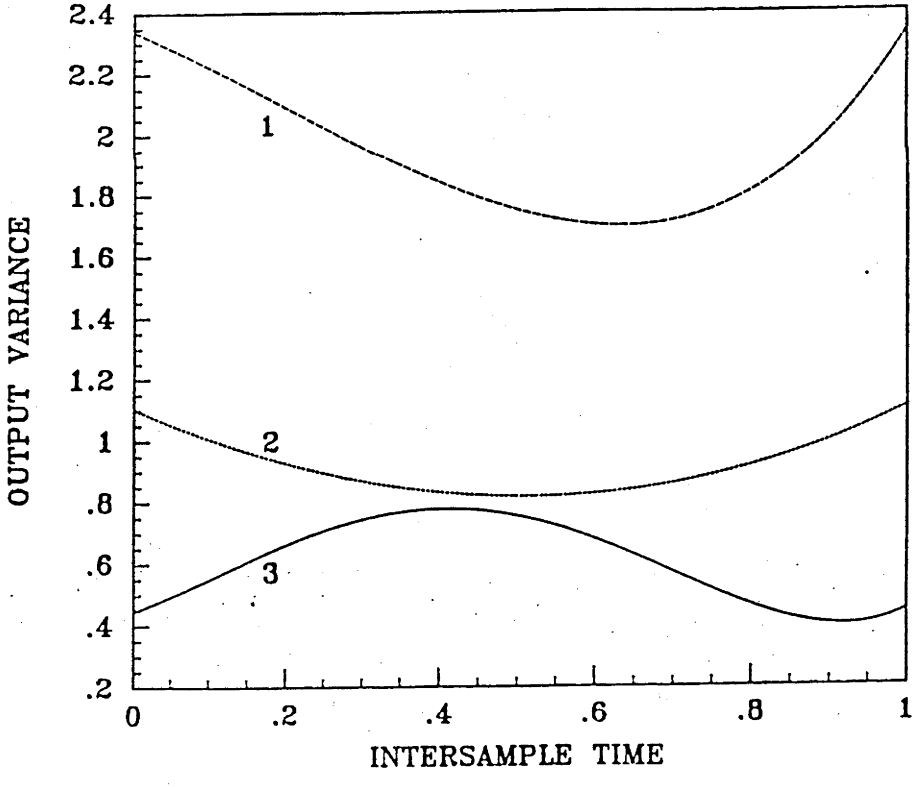


Fig.4.8 The intersample output variance of 2nd order model corresponding to the MVR gains for
1. L = [0.1 1]
2. L = [1 1]
3. L = [1 0.1]

a. $L = [0.1 \quad 1]$; $G_\delta = [0.6667 \quad 1.3333]$

b. $L = [1 \quad 1]$; $G_\delta = [1.6667 \quad 1.8833]$

c. $L = [1 \quad 0.1]$; $G_\delta = [0.0952 \quad 1.0476]$

The intersample variance of output corresponding to the MVR gains were computed and are illustrated in Fig.4.8. For L=[1 0.1], the cheap variance control gain for R=0.015 is G=[1.3433 1.6446]. The effects of increasing R on the normalized maximum of output variance for the cheap control design are shown in Fig.4.5. The optimal R is given by $R^* \approx 0.015$. The effects of increasing R on the normalized maximum of output variance for the MMVR design defined by (4.49a–c) are shown by the curve 2 in Fig.4.6. The optimal R is given by $R^* < 0.1$.

With the initial value $\delta_0 = 0.5$, the worst intersample variance computed in step 2 of the algorithm occurred around $\bar{\sigma} = 0.45$ in all iterations. For fixed $\delta = \bar{\sigma} = 0.45$ the control law $u(kT_c)$ defined by (4.47) is not stabilizing.

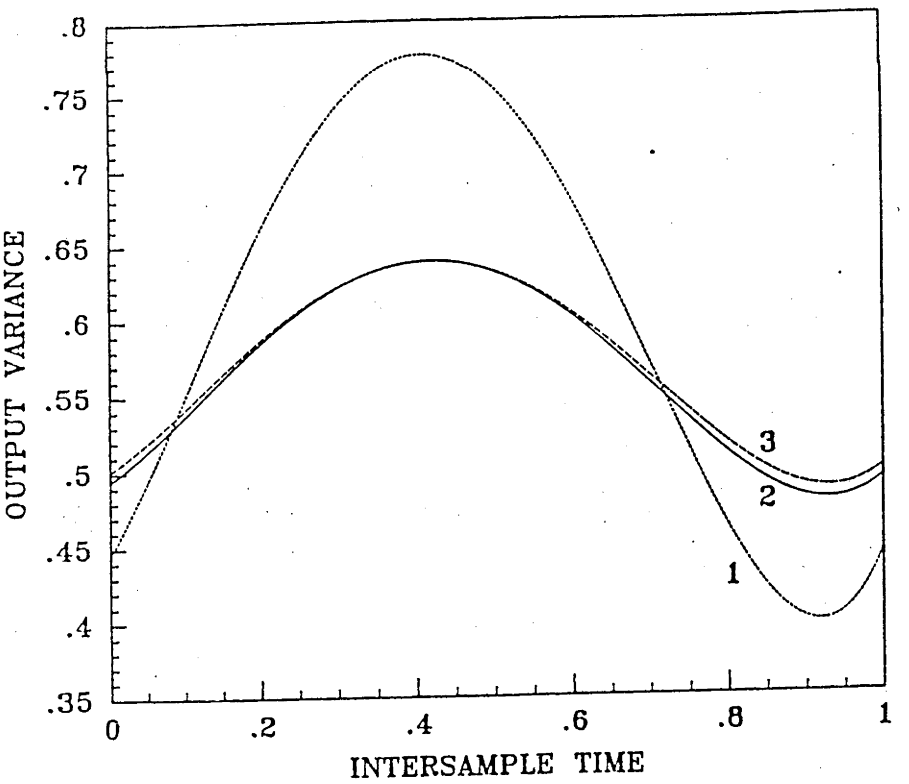


Fig.4.9 The intersample output variance for L=[1 0.1] corresponding to
1. MVR
2. MMVR for $b/a = 10^3$
3. Cheap control for R=0.015

The modified design defined by (4.49a–c) was selected with $b/a=10^3$. The algorithm converged to $G_\delta=[1.3278\ \ 1.6597]$ in 3 iterations. Fig.4.9 shows the intersample variance of output corresponding to the MVR gain (ie. curve 1), the cheap control for R=0.015 (ie. curve 2), and the MMVR gain (ie. curve 3). The effects of increasing the ratio b/a on the normalized maximum of output variance for the MMVR design defined by (4.49a–c) (or by (4.50a–c)) are shown by the curve 2 in Fig.4.7. The optimal ratio b/a is given by $b/a^*>10$. For this example, the intersample cost $J(\sigma,u=G_\delta x)$ is found to be within ±0.01% of its optimum value for $\delta\epsilon[0.3,0.46]$ and $\sigma\epsilon[0.35,0.5]$.

Example 4.4 A sway motion of a ship positioning system described in [Grimble and Patton (1980)] was chosen as a model, and it can be described by a third order plant as in (4.1a–b) with the following parameters

$$A=\begin{bmatrix} -0.0546 & 0 & 0.5435 \\ 1 & 0 & 0 \\ 0 & 0 & -1.55 \end{bmatrix} \quad ; \quad B=\begin{bmatrix} 0 \\ 0 \\ 1.55 \end{bmatrix}$$

$$L'=\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad ; \quad \Omega=\begin{bmatrix} 0.2594 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

With a sampling period $T_c=7$ sec, the discrete model is nonminimum phase. So the MVR design is not suitable for this case and the design should be replaced by a sub–optimal approach. A cheap variance control with a weighting factor R=0.01 was chosen as an output variance regulator. The resulting cheap control law is G=[0.5234 0.0837 0.1706]. The intersample variance behavior of the closed loop system is plotted in Fig.4.10 (ie. curve 1) and the corresponding control variances is $\xi\{u^2(kT_c)\}=5.3$.

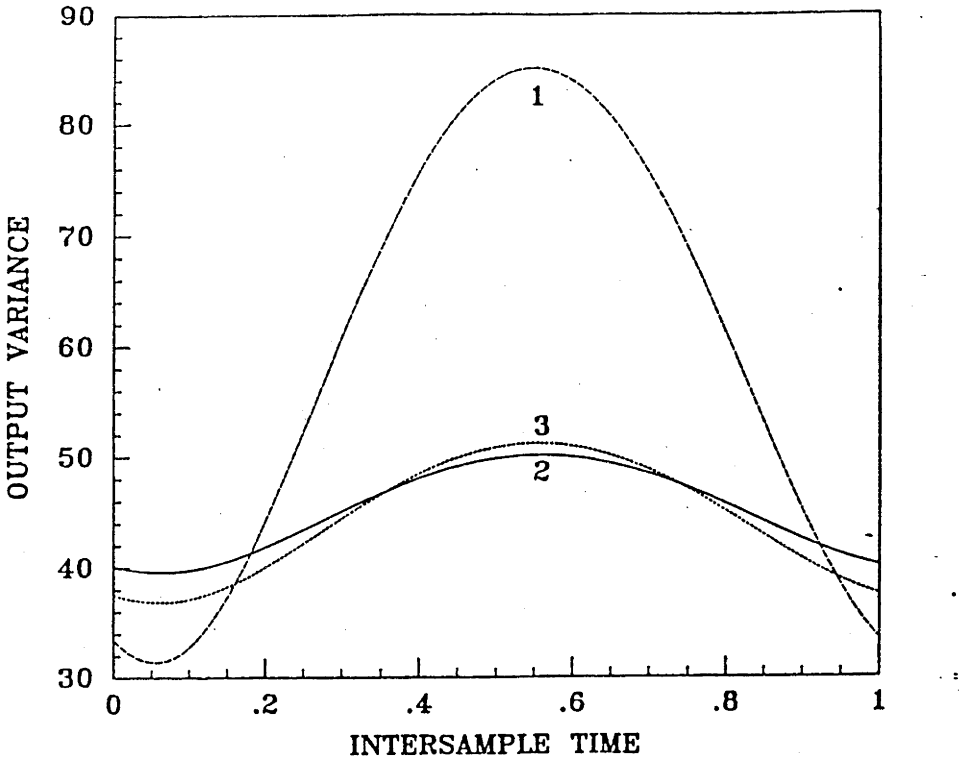


Fig.4.10 The intersample output variance of a 3rd order system corresponding to
1. Cheap control for R=0.01
2. MMVR for $b/a=10^3$
3. KS

For this particular example, a relatively large weighting factor R in the cheap control design was needed to improve the intersample behavior of output. Fig.4.5 (ie. curve 3) illustrates the change in the maximum of output variance as R increases. The optimal R is given by $R^* \approx 10$. The variations of the normalized maximum variance of output as R increases for the MMVR design defined by (4.49a–c) is illustrated by the curve 3 in Fig.4.6. The optimal R is given by $R^* < 0.1$.

The KS method, for this example produces the KS gain G=[0.4615 0.0683 0.1517].

An initial value $\delta_0=0.5$ was selected in the algorithm. The worst intersample cost evaluated at step 2 of the algorithm appeared around $\bar{\sigma}=0.56$ in all iterations. For fixed $\delta=\bar{\sigma}=0.56$ the control law $u(kT_c)$ defined in (4.47) is not stabilizing, so we

chose a ratio $b/a=10^3$ in the MMVR design defined in (4.49a–c). The algorithm converged to the gain G_δ=[0.4375 0.0628 0.1443] in 4 iterations. The effects of increasing the ratio b/a on the maximum of output variance is depicted in Fig.4.7 The optimal ratio b/a is given by $b/a^*>10$. Fig.4.10 shows the variance of outputs in the intersample period corresponding to the cheap control law for R=0.01 (ie. curve 1), to the KS method (ie. curve 2) and to the MMVR design. For this example we found that the intersample cost $J(\sigma,u=G_\delta x)$ is to be within ±0.01% of its optimum value for $\delta \epsilon [0.47,0.6]$ and $\sigma \epsilon [0.5,0.6]$.

4.5 CONCLUSIONS

We have shown that the intersample behavior should be taken into consideration in the design of a digital controller for a continuous plant. In particular, we have illustrated that the minimax output variance regulator (MMVR) can significantly improve the intersample behavior of the closed loop system.

The method of Kwakernaak and Sivan produces a suboptimal design since it minimizes the *average* of output variance in intersample period whereas the MMVR minimizes the *maximum* of output variance. The system performance will be limited by this maximum and not the average.

We have shown that the cheap control design can also be used to minimize the maximum output variance although no procedure is available for choosing the weighting factor R. (It can be seen in Fig.4.5, the factor R is problem dependent). However, for the examples 4.2, 4.3 and 4.4 the suboptimal design of theorem 4.2 produces near optimum performance for

i. $b/a\geq 10$ in (4.49a–c) or (4.50a–c) (see Fig.4.7), or

ii. $R<0.1$ in (4.48a–c) (see Fig.4.6)

For the examples considered in sections 4.4 and 4.5, the algorithm 4.1 converges after 2 updates (<5) of δ_i with δ_0=0.5. It therefore appears that the numerical procedure has the same order of complexity as ordinary (adaptive) LQG design.

Although it is not easy to predict analytically how many local maximums may exist in the intersample period (except for the scalar case), examples (up to 3rd order) show that at most two stationary points (one maximum and one minimum) exist. However, if more than one local maximum occur then the MMVR design still can be accomplished using theorem 4.2 by using several internal weightings corresponding to $N>1$ in (4.19b).

The result that we have developed has assumed that all the plant states are measurable. If we do not have complete knowledge of the states then a conventional state estimator can be incorporated in the design.

CHAPTER 5

FINITE WORDLENGTH LINEAR QUADRATIC REGULATOR DESIGN

5.1 INTRODUCTION

So far we have considered the discrete regulation of continuous time systems under the assumption that the digital control laws can be implemented using infinite precision arithmetic. In the remaining two chapters, we recognize the fact that numbers which may represent compensator input, output, coefficients or states in the (digital) computer must be treated (ie. stored, calculated etc.) with finite accuracy. This so called *finite wordlength* (FWL) problem can be further categorized in two ways, namely *finite state wordlength* (FSWL) and *finite coefficient wordlength* (FCWL) problems.

There are several choices of arithmetics which can be selected to implement the diference equation representing the control algorithm. The most common are *fixed* and *floating point* arithmetics. Floating point arithmetic has relatively large dynamic range compare to fixed point arithmetic but fixed point arithmetic is faster and less expensive to implement. The arithmetic overflow which may occur due to the limited dynamic range of the fixed point arithmetic in many cases can be avoided by scaling the compensator inputs, outputs, coefficients and states. Another advantage of fixed point arithmetic is that the round-off noise analysis is much simpler since the round-off noise is additive [Sripad and Snyder (1977), Barnes et. al (1985)] whereas the round-off noise of floating point arithmetic is multiplicative [Fettweis (1974), Rink and Chong (1979)].

Block floating point arithmetic [Oppenheim (1970), Williamson et. al (1985)] in which only one exponent register is used for all variables is an example of non-standard arithmetic. From computation (ie. additions and multiplications) point

of view, block floating point arithmetic is similar to fixed point arithmetic, and retains some of the advantages of fixed point arithmetic particularly the simplicity of round–off noise analysis. The exponent register of the block floating point arithmetic provides a larger dynamic range than could be achieved by fixed point arithmetic alone. Another example of non–standard arithmetic is the *logarithmic arithmetic* [Kingsbury and Rayner (1971), Lang (1984)]. The main feature of the logarithmic arithmetic is that the quantized values are unevenly spaced. Note that with fixed point numbers, spacing is equal and quantization residue is absolute. With logarithmic number representation, the low magnitude range will have the closest spacing. Consequently, the quantization residue is no longer absolute. High quantization residue is produced in the high magnitude range while more accurate representation (low quantization residue) is achieved in the low magnitude range. In the control applications, during the transient the magnitude of the controller signals (inputs, states or outputs) are often very large in which case a high quantization residue can be tolerated. In steady state operation, the controller signals are generally very low in magnitude, and so a more accurate representation is required. Clearly, the implementation of logarithmic arithmetic would be beneficial. Recently, there has been considerable interest in the logarithmic arithmetic. It has been used in both the digital signal processing [Hall et. al (1970), Kingsbury and Rayner (1971)] and the digital control [Lang (1984), Lamaire and Lang (1986)].

The effects of finite wordlength on the compensator performance are to be analyzed for compensators designed using a linear quadratic Gaussian (LQG) method and implemented using fixed point arithmetic. Finite wordlength representation of the control algorithm produces inaccuracies of compensator coefficients, rounding or truncation after multiplication, and overflow of the additions. We assume throughout that the compensators are properly scaled to avoid overflows during various arithmetic operations. Therefore, the FWL problem that we shall consider is specialized into two problems as follows.

a. The compensator coefficients are represented using an arbitrarily large (but finite) wordlength. Thus, only finite state wordlength (FSWL) problem is considered.

b. The compensator states are realized using an arbitrarily large (but finite) wordlength. Therefore, only finite coefficient wordlength (FCWL) problem is considered.

For recursive algorithms, rounding is required after multiplication in order that the fractional representation does not increase without bound. The cummulative effect of these rounding errors can lead to a significant degradation in the closed-loop performance if a sufficient number of bits is not assigned to the *fractional representation* of compensator states. One solution is to design the compensator assuming it will be implemented using an infinite precision arithmetic (the corresponding performance index will be referred to as the ideal (unquantized state) cost). The fractional wordlength can then be selected so that the resulting degradation in performance is less than a certain percentage of the ideal cost.

Finite precision representation of the compensator coefficients can change the closed loop dynamics considerably. Consequently, the ideal cost is no longer optimal and in extreme cases may not result in a stable closed loop system. A straight-forward way to determine the coefficient wordlength is to design the compensator ignoring the FWL problem. Then, recompute the compensator performance for sets of coefficients that are quantized to different wordlengths, and select the shortest wordlength which gives an 'acceptable' degradation in the compensator performance provided the stability requirement is satisfied.

Different coordinate basis representation do not affect the compensator performance if infinite precision arithmetic is used since all minimal state-space representations of the same transfer function are equivalent. However, the compensator structure plays an important role in the FWL compensator design. The compensator structure is selected such that for a given fractional wordlength (or coefficient wordlength) and for a given compensator gains the compensator

performance is minimized.

With a fixed point implementation, the nodes of the FWL compensator must be properly scaled in order to reduce the possibility of overflow. The overflow problem must be resolved before the FWL performance of different compensator structures can be measured and compared. There are many scaling procedures available which are mainly developed in the digital filter design [Hwang (1975b)]. The most common method is the variance oriented (ie. ℓ_2–scaling) procedure. In digital filter design the ℓ_2–norm of each of the filter state nodes is made equal to the ℓ_2–norm of the filter input. In control applications, the ℓ_2–norm of each of the compensator state nodes and the ℓ_2–norm of the compensator input (or plant output) due only to the external disturbances are equalized.

In [Sripad (1981)], a technique has been presented for analyzing the expected degradation in the performance of a fixed–point arithmetic implementation of a Kalman filter which is originally designed assuming infinite precision implementation. Both FSWL and FCWL problems have been considered. It has been shown that coefficient quantization in the implementation of the filter gains is a significant cause for instability. The effects of state space structures on performance degradation have also been addressed. However, this paper was not concerned with the optimization of the filter structure.

The issues of round–off quantization and scaling in the fixed–point LQG controller implementations have been addressed in [Moroney et. al (1983)]. The LQG design was first carried out ignoring the FWL problem. After scaling has been accomplished, the effects of state quantization noise on control system performance were analysed. Several structures such as cascade, parallel etc. were considered and the respective round–off performance were compared. The optimum structure of [Mullis and Roberts (1976), Hwang (1977)] which minimizes the degradation in LQG cost function from the ideal (ie. unquantized states) was also considered.

In [Moroney et. al (1980)], the coefficient wordlength effects on the performance of the fixed–point LQG compensators were considered. The statistical

wordlength which was first developed in digital filtering [Crochiete (1975)] was used as a measure which reflects a degradation in performance as compared to the ideal (ie. unquantized coefficients) case. Specifically, the statistical wordlength formulation derived from the LQG performance index contains two terms. The first part represents the number of bits necessary to represent the integer portion of the coefficients and the second term provides the number of bits necessary for the fractional portion of the coefficient word. The interesting feature about the statistical wordlength estimate is that it is a continuous function and is differentiable with respect to the coefficients of the structure. Consequently, it can be used as the basis for an iterative (gradient–search constraint) optimization for generating minimum coefficient wordlength structures. However, the scaling issue was not considered in the analysis.

In [Sasahara et. al (1984)], the consequences of finite wordlength implementation of LQG regulators with pre–computed gains were considered. In the round–off noise analysis, a formula which reflects the influence of the compensator structure on the performance is developed using the LQG cost function. The optimum structure of [Mullis and Roberts (1976), Hwang (1977)] has also been considered. The effects of coefficient quantization were shown via the degradation in the performance index from the ideal (ie. unquantized coefficients). Using a deterministic analysis in which the coefficient quantization residues are uniquely determined by the coefficient wordlength and the coefficient matrices, an exact formulation shows the degradation of the LQG cost from the ideal is derived. However, this formula is intractible in the high order case. Furthermore, the calculation requires high precision representation especially to express the coefficient quantization residues. Using a statistical analysis, an attractive formula which gives an approximation of the degradation in the performance from the ideal cost was developed. Simulation studies show that the formula developed using the deterministic analysis gives an accurate result. It is argued that the structure which minimizes the effects of round–off noise (due to the state quantization) also minimizes the effects

of coefficients quantization.

The Kalman filter design which directly takes the finite precision nature of the implementation into account is presented in [Williamson (1985)]. The discrete Kalman filters that are considered are the discrete equivalent of continuous-time systems operating under a fast sampling rate. The optimum FSWL Kalman filter design which includes the state residue feedback compensation is found by means of the standard Riccati technique. It is shown that when the sampling rate is fast and/or the state wordlength is short re-scaling is necessary in order that the ℓ_2-norm at each of the state nodes has similar magnitude. However, the importance of the filter structure was not explored.

In [Williamson (1986)], low noise structures for an N^{th}-order recursive filter incorporating integer residue correction (IRC) [Williamson and Sridharan (1985b)] or sub-optimal error spectrum shaping (ESS) [Abu-El-Haija and Peterson (1979)] has been defined. The IRC term is restricted to be a positive power of two (often $0, \pm 1$) so that it only requires extra additions and shifts. Consequently, the proposed structure only requires N extra additions than would be required by the optimum structure of [Mullis and Roberts (1976), Hwang (1977). The former structure is characterized by a set of filter invariants called the residue modes while the latter is characterized by the singular values (or second order modes). It is shown that the structure proposed in [Williamson (1986] provides a lower round-off noise when the sum of the residue modes is lower than the sum of the second order modes.

In this chapter, we consider both the FSWL and the FCWL problems. The ℓ_2-scaling constraint is used throughout. In the FSWL case, the finite precision nature of the implementation is directly taken into account in selecting the LQG compensator gains. The IRC or the sub-optimal ESS scheme is incorporated in the design. We shall show that the separation theorem [Kwakernaak and Sivan (1972)] which is one of the interesting feature of the LQG design is no longer valid due to existence of state quantization noise. The LQG performance index is used as the basis of an iterative (gradient-search constrained) optimization for generating the

optimal LQG compensator gains. Based on the LQG index, we develop an expression which reflects the influence of the compensator structure on the FSWL performance. The IRC is included in the development. The structure optimization problem is solved by means of the procedures described in [Williamson (1986)] for fixed compensator gains. The optimum solution of the FSWL problem can be achieved by selecting the LQG compensator gains and the compensator structure iteratively.

In the FCWL case, we develop an expression which shows the effects of coefficient quantization on the compensator performance. The expression which is based on the LQG quadratic cost is derived using the statistical approach, and so is only an approximation of the effects of coefficient quantization. Nevertheless, we use the expression to measure and to compare the FCWL performance of different compensator structures. In this case, the ℓ_2–scaling constraint is also used in each structure that we consider.

In section 5.2, we describe a general description of a FWL–LQG regulator. The necessity for scaling and the resulting dynamic range constraint are briefly discussed. In an ideal (ie. infinite precision) situation, the LQG cost function is invariant under a similarity transformation. We show that the performance of a FWL–LQG regulator is strongly affected by the compensator structure. In sections 5.3, 5.4 and 5.5 we investigate the problem of state quantization when the compensator coefficients are implemented using an arbitrarily large (but finite) wordlength. In section 5.3, we do not change the compensator structure. Therefore, the FSWL problem is restricted to finding the optimum Kalman filter and controller gains. In section 5.4, we assume the compensator gains are fixed. The FSWL problem is then to optimize the LQG cost function with respect to the compensator structure. The optimum FSWL solution which can be derived by iteratively solving the FSWL problem for fixed structure and the FSWL problem for fixed gains is described in section 5.5. In section 5.6, we consider the effects of coefficient quantization on the compensator performance. First, we assume that the effects of state quantization are negligible. Using a

statistical analysis, we develop an expression based on the LQG cost function which can be used to measure the consequences of using short coefficient wordlength. Illustrative examples are included throughout.

5.2 FINITE WORDLENGTH LINEAR QUADRATIC GAUSSIAN REGULATOR

In this section we discuss the effects of the finite register length on the design of infinite horizon linear quadratic Gaussian (LQG) regulators. Specifically, consider the minimal discrete-time model described by

$$x(k+1) = \Phi x(k) + \Gamma u(k) + \omega(k) \tag{5.1a}$$

$$y(k) = Lx(k) + \eta(k) \tag{5.1b}$$

where $x(k)$ is a simplified notation of $x(kT_c)$ where T_c is a sampling period. The dimensions of the state, the input and the output are

$$x(t)\epsilon R^{n_x} \quad ; \quad u(t)\epsilon R^{n_u} \quad ; \quad y(t)\epsilon R^{n_y}$$

The processes $\{\omega(k)\}$ and $\{\eta(k)\}$ are zero-mean wide-sense stationary (WSS) processes having covariance

$$\xi\{\begin{bmatrix}\omega(k)\\ \eta(k)\end{bmatrix}[\omega'(k)\ \eta'(k)]\} = \begin{bmatrix}\Omega & 0\\ 0 & \Lambda\end{bmatrix} \tag{5.2}$$

The infinite horizon quadratic cost which we seek to minimize is described by

$$J = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-\infty}^{m} [x'(k)Qx(k)+2x'(k)Mu(k)+u'(k)Ru(k)]\} \tag{5.3}$$

where the weighting matrices $Q\geq 0$, M and $R>0$ are respectively $(n_x x n_x)$, $(n_x x n_u)$ and $(n_u x n_u)$. In an ideal situation (ie. infinite wordlength) the optimal compensator which minimizes the performance index (5.3) subject to the plant (5.1a-b) is governed by

$$\hat{x}(k+1) = \Phi\hat{x}(k) + \Gamma u(k) + K(y(k)-L\hat{x}(k)) \tag{5.4a}$$

$$u(k) = - G\hat{x}(k) \tag{5.4b}$$

where the predicted state $\hat{x}(k)$ is a simplified notation of $\hat{x}(k|y(k-1))$ and where the

matrices K and G are respectively the Kalman filter and the controller gains. Define the prediction error

$$\varepsilon(k) \triangleq x(k) - \hat{x}(k) \tag{5.5a}$$

From (5.1a–b) and (5.4a–b), we obtain the augmented representation

$$\begin{bmatrix}\varepsilon(k+1)\\ \hat{x}(k+1)\end{bmatrix} = \begin{bmatrix}\Phi-KL & 0\\ KL & \Phi-\Gamma G\end{bmatrix}\begin{bmatrix}\varepsilon(k)\\ \hat{x}(k)\end{bmatrix} + \begin{bmatrix}I & -K\\ 0 & K\end{bmatrix}\begin{bmatrix}\omega(k)\\ \eta(k)\end{bmatrix} \tag{5.5b}$$

Define the covariance matrix

$$S \triangleq \xi\{\varepsilon(k)\varepsilon'(k)\} \tag{5.5c}$$

The optimal Kalman filter and the controller gains K and G are given by (as in lemma 4.2 of the previous chapter)

$$K = \Phi S L'(LSL'+\Lambda)^{-1} \tag{5.6a}$$

$$\bar{G} = (R+\Gamma'\Sigma\Gamma)^{-1}\Gamma'\Sigma\bar{\Phi} \tag{5.6b}$$

where the matrices S and Σ satisfy the algebraic Riccati equations

$$S = (\Phi-KL)S(\Phi-KL)' + \Omega + K\Lambda K' \tag{5.6c}$$

$$\Sigma = (\bar{\Phi}-\Gamma\bar{G})'\Sigma(\bar{\Phi}-\Gamma\bar{G}) + \bar{Q} + \bar{G}R\bar{G}' \tag{5.6d}$$

where the matrices $\bar{\Phi}$, $\bar{G}$ and $\bar{Q}$ are given by

$$\bar{\Phi} = \Phi-\Gamma R^{-1}M' \tag{5.6e}$$

$$\bar{Q} = Q-MR^{-1}M' \tag{5.6f}$$

$$\bar{G} = G-R^{-1}M' \tag{5.6g}$$

By definition [Chan et. al (1984)], the gains K and G are stabilizing (ie. when $|\lambda_i(\Phi-KL)|<1$ and $|\lambda_i(\Phi-\Gamma G)|<1$ for all $i\epsilon[1,n_x]$) when the matrices S and Σ are the stabilizing solutions (which symmetric and positive definite) of (5.6c–d). The stabilizing solutions S and Σ exist, and unique (ie. unique P and unique Σ) if and only if

(i) $\{\Phi,D\}$ has no uncontrollable modes on the unit circle and

(ii) $\{\bar{\Phi},\bar{D}\}$ has no unobservable modes on the unit circle

where $\Omega=DD'$ and $\bar{Q}=\bar{D}\bar{D}'$. We assume both the Kalman filter and the controller gains K and G are *stabilizing*. The implementation of a digital compensator is shown in Fig.5.1. In Fig.5.1, it can be seen that there are two interfaces; they are

the digital converter system which includes a digital-to-analog converter (DAC) and the analog acquisition system which contains an analog-to-digital (ADC) converter. Thus, the finite wordlength implementation should also consider the implications of both the DAC and the ADC. However, we will not discuss the finite wordlength implementation of these interfaces except when we discuss the scaling issue.

In the real situation, an LQG compensator is implemented using a finite wordlength arithmetic. We assume throughout that the LQG compensator is realized using the *fixed-point* arithmetic. Therefore, we only concern with the fixed-point round-off and coefficient quantization residues.

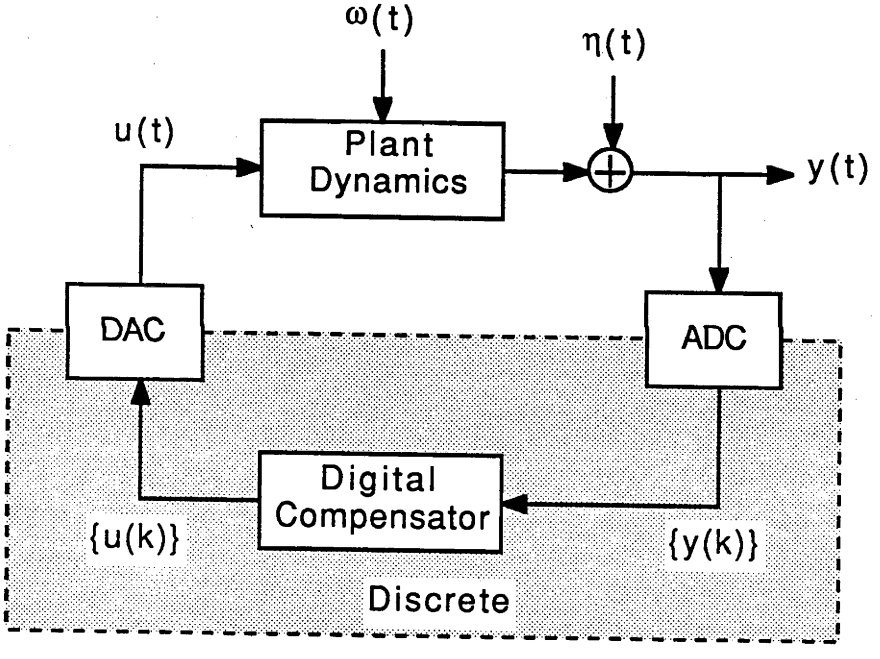


Fig.5.1 Digital control system configuration

The fixed-point finite wordlength (FWL) version of the ideal LQG compensator (5.4a-b) can be described as

$$\hat{x}^*(k+1) = \Phi^* Q[\hat{x}^*(k)] + \Gamma^* u^*(k) + K^*(y^*(k) - L^* Q[\hat{x}^*(k)]) \qquad (5.7a)$$

$$u^*(k) = - Q[G^* Q[\hat{x}^*(k)]] \qquad (5.7b)$$

where the coefficient matrix A^* denotes the finite coefficient wordlength (FCWL) representation of A and satisfies

$$A^* = A + \Delta A \tag{5.7c}$$

and where $Q[\hat{x}^*(k)]$ represents the quantized value of the predicted state $\hat{x}^*(k)$ and satisfies

$$Q[\hat{x}^*(k)] = \hat{x}^*(k) - e(k) \tag{5.7d}$$

The properties of the residues ΔA and $e(k)$ will be discussed later in the chapter.

The discrete plant response when controlled by the compensator (5.7a–b) can be expressed as

$$x^*(k+1) = \Phi x^*(k) + \Gamma u^*(k) + \omega(k) \tag{5.8a}$$

$$y^*(k) = Lx^*(k) + \eta(k) \tag{5.8b}$$

Note that the states $x^*(k)$, the input $u^*(k)$ and the output $y^*(k)$ are not the same with $x(k)$, $u(k)$ and $y(k)$ in (5.1a–b) due to FWL implementation in (5.7). Consequently, the optimized performance index differs from the ideal cost (5.3), and is described by

$$J^* = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} [\,(x^*(k))'Qx^*(k) + 2(x^*(k))'Mu^*(k) + (u^*(k))'Ru^*(k)\,]\} \tag{5.9a}$$

Define the difference between the FWL cost J^* and the ideal cost J as follows

$$\Delta J \triangleq J^* - J \tag{5.9b}$$

The scalar ΔJ can be used as a degradation metric of the compensator performance due to the FWL implementation.

ℓ_2–scaling : The possibility of an (arithmetic) overflow needs to be reduced to an 'acceptable' degree particularly when the compensator is implemented using a fixed–point arithmetic. This requirement can be met by scaling the compensator variables and parameters. However, scaling is not merely meant for decreasing the possibility of overflow, it is also aimed at altering the compensator coefficients such that they are all lie within a prescribed dynamic range. It is well known that the latter objective is not always achieved if the scaling is only oriented towards overflow reduction. The scaling task can be divided into three different categories;

namely the *input/output scaling*, the *state scaling* and the *coefficient scaling*. In this chapter, we assume the LQG compensators are implemented using pre-scaled coefficients. Therefore, the scaling issue that we consider concerns only the input/output and the state scaling. Before we discuss the input/output and the state scaling procedures let consider the scaling unit or measure. Basically, the dynamic range constraint can be specified using either a deterministic [Jackson (1970a)] or a statistical [Hwang (1975)] approach. These two approaches are discussed in some depth in [Moroney (1983)] and it has been shown that the variance oriented (ie. ℓ_2) scaling constraint is the most convenient to use. In fact, this type of scaling has been widely used in control applications [Moroney et. al (1980,1983), Sasahara et. al (1984), Scharf and Sigurdson (1984), Ahmed and Belanger (1984)]. The ℓ_2-norm for the (infinite) sequence $\{a(k)\}$ for $k\epsilon[0,\infty)$ is defined as [Epstein (1970)]

$$\|a(k)\|_2 \triangleq \left\{ \sum_{k=0}^{\infty} |a(k)|^2 \right\}^{\frac{1}{2}}$$

There are many considerations that contribute to the selection of scaling factors. In [Moroney (1983)], the considered compensator representation was rearranged such that it has only one (scalar) input y(t). Therefore, the ℓ_2-scaling procedure developed for digital filters can be adopted; that is, the ℓ_2-scaling constraint is employed to make the variance of each state of the compensator equals to the variance of the compensator input which has been normalized to unity. Therefore, the ℓ_2-scaling constraint is

$$\|\hat{z}_j^*(k)\|_2 = 1$$

for all $j\epsilon[1,n_z]$ where $n_z(=n_x)$ is the dimension of the scaled state $z^*(k)$ discussed below. In the sequel, we term this the *unity* ℓ_2*-scaling* constraint. This approach however ignores additional scaling that is generally required in order to handle initial condition type disturbance transients for which ℓ_∞-scaling is more appropriate. Our approach is therefore to compute the FWL-LQG design based on some given ℓ_2-scaling of the compensator states without attempting to justify a correct level of scaling. In practice, simulation studies would most likely aid this process.

The FWL compensator (5.7a) can be represented (and hence implemented) in different ways. Three different representations are listed as follows.

$$\hat{x}^*(k+1) = (\Phi-\Gamma G-KL)^*Q[\hat{x}^*(k)] + K^*y^*(k) \qquad (5.10a)$$

$$\hat{x}^*(k+1) = (\Phi-KL)^*Q[\hat{x}^*(k)] + \Gamma^*u^*(k) + K^*y^*(k) \qquad (5.10b)$$

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)^*Q[\hat{x}^*(k)] + K^*y_e^*(k) \qquad (5.10c)$$

where the FCWL matrix A^* is defined by (5.7c), $Q[\hat{x}^*(k)]$ is defined by (5.7d) and where

$$y_e^*(k) = y^*(k) - L^*Q[\hat{x}^*(k)] \qquad (5.10d)$$

In this chapter, we assume that the plants (and hence the compensators) are single-input and single-output (SISO) plants. We shall show that this assumption is made mainly because the scaling procedures for the SISO compensators are different to the scaling procedures of the multi-input multi-output (MIMO) compensator. Therefore, the input u(t) and the output y(t) throughout this chapter are single variables; that is

$$n_u = n_y = 1$$

The compensator (5.10a) is the same with the compensator (5.10c) in that they both have only one input. However, the compensator (5.10a) possesses two drawbacks. First, the compensator input (or plant output) $y^*(k)$ is not 'white' even though the process disturbance $\{\omega(k)\}$ and the measurement noise $\{\eta(k)\}$ in (5.8a-b) are both 'white'. Secondly, the compensator (5.10a) may not be (open-loop) stable. Therefore, the ℓ_2-scaling objective can not be met.

Recall that the Kalman filter and the controller gain K and G are assumed to be stabilizing (ie. $\Phi-KL$ and $\Phi-\Gamma G$ are both stables). Therefore, from the (open-loop) stability point of view, the compensators (5.10b-c) are preferred. However, the compensator (5.10b) has two inputs (ie. $u^*(k)$ and $y^*(k)$). The variance of the input $u^*(k)$ and the input $y^*(k)$ are not the same. Thus, we can note equalize the probabilities of overflow at every state and every input. One solution is to make the variance (or the probability of overflow) of each state equals to the variance of one of the compensator input; that is, the input with the larger

variance. In fact, this is the scaling procedure that we will use for MIMO compensators which will be examined in section 6.3 of the next chapter. Compensator (5.10b) is preferred if the realization $\{\Phi,L\}$ is in an observable canonical form [Kailath (1980)] since the realization $\{\Phi-KL,L\}$ is also in an observable canonical form.

Compensator (5.10c) is preferred if the realization $\{\Phi,\Gamma\}$ is in a controllable canonical form [Kailath (1980)] since the realization $\{\Phi-\Gamma G,\Gamma\}$ is also in a controllable canonical form. Another advantage of the compensator (5.10c) is that the input $y_e^*(k)$ defined by (5.10d) is well known to be 'white'. This is the so called the innovations sequence [Anderson and Moore (1979)]. In this chapter, the FWL analysis will be based on the compensator (5.10c). The FWL compensator (5.10c) can be represented using a different coordinate basis. Let us apply a similarity transformation T (where T^{-1} exists) and consider the compensator with the transformed structure as defined by

$$\hat{z}^*(k+1) = \tilde{\Phi}_2^* Q[\hat{z}^*(k)] + \tilde{K}^* \tilde{y}_e^*(k) \tag{5.11a}$$

$$\tilde{y}_e^*(k) = y^*(k) - \tilde{L}^* Q[\hat{z}^*(k)] \tag{5.11b}$$

$$Q[\hat{z}^*(k)] = \hat{z}^*(k) - \epsilon(k) \tag{5.11c}$$

where

$$\tilde{\Phi}_2^* = (T^{-1}(\Phi-\Gamma G)T)^* \tag{5.11d}$$

$$\tilde{K}^* = (T^{-1}K)^* \quad ; \quad \tilde{L}^* = (LT)^* \tag{5.11e}$$

Replacing $x^*(k)$ in (5.9a) by $z^*(k)$ gives the following transformed quadratic cost.

$$\tilde{J}^* = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} [(z^*(k)' \tilde{Q} z^*(k) + 2z^*(k)' \tilde{M} u^*(k) + u^*(k)' R u^*(k)]\} \tag{5.12a}$$

where the weighting matrices $\tilde{Q}$ and $\tilde{M}$ are related to the weighting matrices Q and M in (5.9a) by

$$\tilde{Q} = T' Q T \tag{5.12b}$$

$$\tilde{M} = T' M \tag{5.12c}$$

The state quantization residue $\epsilon(k)$ will be examined in the later sections.

State vector scaling, as discussed in [Mullis and Roberts (1976b)] corresponds to a diagonal equivalent transformation of the unscaled structure in (5.10c). From the transformed representation (5.11a–c) and the realization (5.10c), the state scaling can be defined by

$$\hat{z}^*(k) = S^{-1}\hat{x}^*(k) \tag{5.13b}$$

where the scaling matrix S is diagonal, and is given by

$$S = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{n_x} \end{bmatrix} \tag{5.13c}$$

This diagonal scaling certainly affects the compensator performance (ie. $J^* \neq \tilde{J}^*$ in (5.9a) and (5.12a)). The effects of the state and the input/output scalings on the compensator performance will be investigated in section 5.3.

To use the unity ℓ_2–scaling constraint, the compensator input has to be normalized to unity. This can be achieved by multiplying the compensator input by a scaling factor s which is defined by

$$s = \frac{1}{(\sigma_y^2)^{\frac{1}{2}}} \tag{5.14a}$$

where

$$\sigma_y^2 \triangleq \xi\{\tilde{y}_e^*(k)^2\} \tag{5.14b}$$

where $\tilde{y}_e^*(k)$ is defined in (5.11b). This scaling factor may be incorporated in the ADC. In order to preserve the closed-loop transfer function, the output of the compensator has to be multiplied by s^{-1} where s is defined by (5.14a). Similar to the input case, the output scaling factor can be incorporated in the DAC. In terms of a similarity transformation, the input/output scaling is equivalent to making

$$\hat{z}^*(k) = \frac{\hat{x}^*(k)}{s} \tag{5.14c}$$

where s is a (non zero) scalar (and may be given by (5.14a)).

5.3 OPTIMUM FINITE STATE WORDLENGTH GAINS : DEFAULT STRUCTURE

In this section, we examine a simplified version of the FWL problem discussed in the previous section. That is, we assume that the coefficient quantization residues are negligible compared to the state quantization (or round-off) residues. This simplified problem is called the *finite state wordlength* (FSWL) problem. In this context, the FSWL compensator can be represented as a simplified version of the FWL compensator (5.10c); that is

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)Q[\hat{x}^*(k)] + Ky_e^*(k) \tag{5.15a}$$

$$Q[\hat{x}^*(k)] = \hat{x}^*(k) - e(k) \tag{5.15b}$$

$$u^*(k) = - Q[GQ[\hat{x}^*(k)]] = - GQ[\hat{x}^*(k)] + d(k) \tag{5.15c}$$

$$y_e^*(k) = y^*(k) - LQ[\hat{x}^*(k)] \tag{5.15d}$$

The compensator state $Q[\hat{x}^*(k)]$ will be shown under appropriate conditions to provide an estimate of the state $x(k)$ of the physical model as defined by (5.1a-b). This compensator is therefore said to have the *default structure*. Any new state space structure resulting from the transformation

$$\hat{z}^*(k) = T^{-1}\hat{x}^*(k) \tag{5.16}$$

where T is *diagonal* will also qualify as a default structure since components of $\hat{z}(k)$ will then be proportional to the estimate of the physical states. Later in section 5.4, we shall consider the choice of the optimal state space structure in which T in (5.16) is permitted greater freedom.

The components of $Q[\hat{x}^*(k)]$ are assumed to be less than unity in magnitude with a B-bit fractional representation while the compensator coefficients Φ-KL, Γ, K and G while not necessarily less than unity all have an exact fractional B_c-bit representation. In this section, we only examine the effects of the fractional wordlength B on the compensator performance, and not the effects of coefficient quantization. Hence, we shall assume B_c is arbitrary large but finite.

Note that double precision additions (of fractional width $B+B_c$) is implied in (5.15a) so that $\hat{x}^*(k+1)$ has a $B+B_c$ bit fractional representation. However, only

$Q[\hat{x}^*(k+1)]$ is stored for subsequent calculations. Note also that $GQ[\hat{x}^*(k)]$ has a $B+B_c$ bit fraction so that an *additional* quantization back to B bits is required so as to produce the correct wordlength for u(k) suitable for DAC operation.

With a fixed point implementation, the nodes of the FSWL compensator must be properly scaled in order to reduce the possibility of overflow to an 'acceptable' degree. Assuming no overflows occurs, each component $e_n(k)$ of the quantization residue e(k) in (5.15b) obtained by rounding satisfies $|e_n(k)| < 2^{-(B+1)}$ [Phillips and Nagle (1984)]. Furthermore, due to the presence of process and measurement noise and assuming satisfactory scaling of the compensator states we are justified [Sripad and Snyder (1977), Barnes et. al (1985)] in modelling the quantization residue {e(k)} in (5.15b) as a zero mean 'white noise' process with covariance matrix qI_x where

$$q = \frac{1}{12} 2^{-2B} \tag{5.17}$$

and where I_x is an $(n_x x n_x)$ identity matrix where n_x is the dimension of the state x(k). The quantization of the compensator input $u^*(k)$ produces the round-off error d(k) in (5.15c) which can also be assumed to be zero mean 'white' noise with covariance qI_u (I_u is an $n_u x n_u$ identity matrix where n_u is the dimension of the control u(k)).

After substitution of $Q[\hat{x}^*(k)]$ in (5.15a) and (5.15c) using (5.15b), the FSWL compensator realization can be described by

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)(\hat{x}^*(k)-e(k)) + Ky_e^*(k) \tag{5.18a}$$

$$u^*(k) = -G\hat{x}^*(k) + Ge(k) + d(k) \tag{5.18b}$$

where $y_e^*(k)$ is defined in (5.15d). A technique known as the *integer residue feedback* (IRC) or the sub-optimal *error spectrum shaping* (ESS) [Abu-EL-Haija and Peterson (1979), Munson and Liu (1981), Williamson and Sridharan (1985b)] has been used extensively in the filter design as a mean of reducing the effects of state quantization on the filter performance. The implementation of this IRC approach on the FSWL compensator (5.18a) gives

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)(\hat{x}^*(k)-e(k)) + \varphi e(k) + Ky_e^*(k) \tag{5.19}$$

where φ is an $(n_x x n_x)$ matrix whose components are *integers*, so that wordlength consistency is preserved (ie. both the left and the right hand sides of (5.19) can be exactly represented using $B+B_c$ bit fractions).

We shall see later that the minimum cost J^* described in the previous section is achieved by the optimum selection of the transformation matrix T and of the integer residue matrix φ. Unfortunately, the general problem of finding the optimal integer residue matrix φ and the optimal transformation matrix T does not have a closed-form solution. To overcome this difficulty, we then seek the sub-optimal solution by restricting φ to I_x or $-I_x$ depending on the closed loop system matrix Φ-KL [Williamson (1986)]. In the following example we show the effects of the integer residue correction on the performance index J^*.

Example 5.1 Consider the minimal 6^{th}-order SISO model (5.8a-b) where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively.

Wordlength B	$J^*(x10^{-5})$	
	φ=I	φ=0 (without IRC)
14	9.6029	9.6029
12	9.6030	9.6035
10	9.6276	9.6301
8	9.7667	21.0268
6	10.1372	#
4	16.2136	#

Table 5.1. The effects of the integer residue correction (IRC) on the compensator performance.

Consider also the quadratic cost J^* defined by (5.9a) where the weighting factors Q, M and R are respectively given by (A.23), (A.24) and (A.25).

The ideal (ie. infinite precision) Kalman filter and the controller gains K_∞ and G_∞ which can be found by solving the algebraic Riccati equations (5.6c-d) are given by (A.28) and (A.29). Table 5.1 shows the cost J^* which were computed using the procedures developed later (see lemma 5.4) for different values of fractional wordlength B and for φ=0 (ie. without residue correction) and for $\varphi=I_x$ (where I_x is an $(n_x x n_x)$ identity matrix) subject to the unity ℓ_2-scaling constraint. The notation # in table 5.1 means the compensator states can not be scaled to satisfy the unity ℓ_2-scaling constraint due to the contribution of the FSWL noise which makes the variance of each state of the compensator larger than unity.

□□□

Define the FSWL state prediction error

$$\varepsilon^*(k) \triangleq x^*(k) - \hat{x}^*(k) \tag{5.20a}$$

Then from (5.8a) and (5.19), we obtain the FSWL prediction error equation which is governed by

$$\varepsilon^*(k+1) = (\Phi-KL)(\varepsilon^*(k) + e(k)) + \omega(k) - K\eta(k) - \varphi e(k) \tag{5.20b}$$

From (5.19) and (5.20b), we obtain an augmented representation

$$\begin{bmatrix} \varepsilon^*(k+1) \\ \hat{x}(k+1) \end{bmatrix} = \Phi a \begin{bmatrix} \varepsilon^*(k)+e(k) \\ \hat{x}(k)-e(k) \end{bmatrix} + K_a \begin{bmatrix} \omega(k) \\ \eta(k) \end{bmatrix} + \begin{bmatrix} \Gamma \\ 0 \end{bmatrix} d(k) - \begin{bmatrix} \varphi & 0 \\ 0 & \varphi \end{bmatrix} \begin{bmatrix} e(k) \\ -e(k) \end{bmatrix} \tag{5.21a}$$

where

$$\Phi_a = \begin{bmatrix} \Phi-KL & 0 \\ KL & \Phi-\Gamma G \end{bmatrix} \tag{5.21b}$$

$$K_a = \begin{bmatrix} I_x & -K \\ 0 & K \end{bmatrix} \tag{5.21c}$$

Note that the control vector $u^*(k)$ in (5.18b) has been used in deriving the augmented representation (5.21a). The quadratic cost (5.9a) is measured in terms of

the plant state $x^*(k)$. For convenience, the quadratic cost J^* was modified so now it is expressed in terms of the state estimation error $\epsilon^*(k)$ defined in (5.20a) and the predicted state $\hat{x}^*(k)$. Substitution of $x^*(k)$ in (5.9a) using (5.20a) and substitution of $u^*(k)$ in (5.9a) using (5.18b) yields

$$J^* = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} s(k)\} \tag{5.22a}$$

where the function s(k) is given by

$$s(k) = \epsilon^*(k)'Q\epsilon^*(k) + 2\epsilon^*(k)'(Q-MG)\hat{x}^*(k) +$$

$$\hat{x}^*(k)'(Q-2MG+G'RG)\hat{x}^*(k) + e'(k)G'RGe(k) + d'(k)Rd(k) \tag{5.22b}$$

From the augmented representation (5.21a–c), we have in steady state that

$$\xi\left\{\begin{bmatrix}\epsilon^*(k)\\ \hat{x}^*(k)\end{bmatrix}\begin{bmatrix}\epsilon^*(k)\\ \hat{x}^*(k)\end{bmatrix}'\right\} \triangleq P = P_1 + qP_2 \tag{5.23a}$$

where q is defined in (5.17) and where the covariance matrices P_1 and P_2 satisfy the following Lyapunov equations

$$P_1 = \Phi_a P_1 \Phi_a' + K_a \Theta K_a' + qHH' \tag{5.23b}$$

$$P_2 = \Phi_a P_2 \Phi_a' + (\Phi_a - \varphi_a) I_a (\Phi_a - \varphi_a)' \tag{5.23c}$$

where Φ_a and K_a are defined in (5.21b–c) and where

$$\varphi_a = \begin{bmatrix}\varphi & 0\\ 0 & \varphi\end{bmatrix} \quad ; \quad I_a = \begin{bmatrix}I_x & -I_x\\ -I_x & I_x\end{bmatrix} \tag{5.23d}$$

$$H = \begin{bmatrix}\Gamma\\ 0\end{bmatrix} \quad ; \quad \Theta = \begin{bmatrix}\Omega & 0\\ 0 & \Lambda\end{bmatrix} \tag{5.23e}$$

For a given set of compensator gains K and G and the fractional wordlength B, the performance of the FSWL compensator (5.19) is indicated by J^* in (5.22a) which can be calculated statistically from the augmented equation (5.21a–c). Using the covariance equations (5.23b–c), the computation of the performance index J^* can be simplified using a formulae that we present in the following lemma.

Lemma 5.1 Consider the finite state wordlength compensator (5.19) and the corresponding performance index (5.22a–b). Consider as well the augmented representation (5.21a–c) and the covariance equations (5.23b–c). Then, the quadratic cost J^* in (5.22a–b) can be rewritten as

$$J^* = tr(\Upsilon P_1) + q[tr(\Upsilon P_2) + tr(R)] \qquad (5.24a)$$

where tr(Z) denotes the trace of a square matrix Z and where the square matrix Υ is given by

$$\Upsilon = \begin{bmatrix} Q & Q-MG \\ (Q-MG)' & Q-2MG+G'RG \end{bmatrix} \qquad (5.24b)$$

Proof : The function s(k) in (5.22b) can be rearranged into the following representation

$$s(k) = [\epsilon^*(k)' \ \hat{x}^*(k)']\Upsilon[\epsilon^*(k)' \ \hat{x}^*(k)']' + e'(k)[I_X \ -I_X]\Upsilon[I_X \ -I_X]'e(k) + d'(k)Rd(k) \qquad (5.25)$$

where Υ is given by (5.24b). Substitute s(k) in (5.22a) using (5.25), use the fact that for a non–zero vector g and a positive semi–definite matrix V

$$\xi\{g'Vg\} = tr[V\xi\{gg'\}]$$

and utilize the covariance matrices P_1 and P_2 defined in (5.23a–c) to get the result in (5.24a).

□□□

In section 5.2, we briefly reviewed the derivation of an ideal LQG regulator. Using lemma 5.1 above, the ideal LQG design can be reformulated as follows.

Lemma 5.2 Consider the discrete model described by (5.1a–b) and the quadratic cost (5.3). Assume $\{\Phi,\Gamma,L\}$ is minimal and also assume that the pairs $\{\Phi,D\}$ and $\{\Phi,\bar{D}\}$ are respectively controllable and observable where

$$DD' = \Omega \quad ; \quad \bar{D}\bar{D}' = \bar{Q} \qquad (5.26a)$$

where

$$\bar{Q} = Q - MR^{-1}M' \qquad (5.26b)$$

Then, the optimal steady state Kalman filter gain K_∞ and the optimal steady state controller gain G_∞ which minimize the performance index (5.3) are given by

$$K_\infty = \Phi P_\infty(1,1) L' (L P_\infty L' + \Lambda)^{-1} \tag{5.27a}$$

$$\bar{G}_\infty = (R + \Gamma' \Sigma \Gamma)^{-1} \Gamma' \bar{\Sigma} \bar{\Phi} \tag{5.27b}$$

where the matrices $P_\infty(1,1)$ and Σ are the stabilizing solution of the following algebraic Riccati equations

$$P_\infty(1,1) = (\Phi - KL) P_\infty(1,1) (\Phi - KL)' + \Omega + K_\infty \Lambda K_\infty' \tag{5.28a}$$

$$\Sigma = (\bar{\Phi} - \Gamma \bar{G}_\infty)' \Sigma (\bar{\Phi} - \Gamma \bar{G}_\infty) + \bar{Q} + \bar{G}_\infty' R \bar{G}_\infty \tag{5.28b}$$

where

$$\bar{\Phi} = \Phi - \Gamma R^{-1} M' \tag{5.28c}$$

$$\bar{G}_\infty = G_\infty - R^{-1} M \tag{5.28d}$$

Moreover, the minimal cost is given by

$$tr(\Upsilon P_\infty) = tr(Q P_\infty(1,1)) + tr(K_\infty (L P_\infty(1,1) L' + \Lambda) K_\infty' \Sigma) \tag{5.29}$$

where P_∞ is the equivalent of P defined in (5.23a-c) for q=0 (or B=∞ in (5.17)).

Proof : The derivation of the optimal gains K_∞ and $\bar{G}_\infty$ in (5.27a-b) and the algebraic Riccati equations (5.28a-b) can be found in [Kwakernaak and Sivan (1972)]. Let P_∞ is partitioned as

$$P_\infty = \begin{bmatrix} P_\infty(1,1) & P_\infty(1,2) \\ P_\infty(2,1) & P_\infty(2,2) \end{bmatrix} \tag{5.30a}$$

By the separation theorem discussed in chapter 3 the sub-matrices

$$P_\infty(1,2) = P_\infty(2,1) = 0 \tag{5.30b}$$

Using (5.30b) and from (5.24a-b) we obtain

$$tr(\Upsilon P_\infty) = tr\left[\begin{bmatrix} Q & Q - MG_\infty \\ (Q - MG_\infty)' & Q - 2MG_\infty + G_\infty' R G_\infty \end{bmatrix} \begin{bmatrix} P_\infty(1,1) & 0 \\ 0 & P_\infty(2,2) \end{bmatrix}\right]$$

$$= tr(Q P_\infty(1,1))) + tr((Q - 2MG_\infty + G_\infty' R G_\infty) P_\infty(2,2)) \tag{5.31}$$

where $P_\infty(2,2)$ can be derived from (5.21a) and is given by

$$P_\infty(2,2) = (\Phi - \Gamma G_\infty) P_\infty(2,2) (\Phi - \Gamma G_\infty)' + K_\infty (L P_\infty(1,1) L' + \Lambda) K_\infty' \tag{5.32a}$$

Substitution of $\bar{\Phi}$, $\bar{G}_\infty$ and $\bar{Q}$ in (5.28b) using (5.28c-d) and (5.26b) respectively yields

$$\Sigma = (\Phi-\Gamma G_\infty)'\Sigma(\Phi-\Gamma G_\infty) + Q - 2MG_\infty + G_\infty'RG_\infty \qquad (5.32b)$$

The following is a property of the trace operator

$$tr((A-HAH')B) = tr(A(B-H'BH)) \qquad (5.32c)$$

Letting

$$A = P_\infty(2,2)$$

$$B = \Sigma$$

$$H = \Phi - \Gamma G_\infty$$

and using (5.32c), the minimum cost (5.31) can be rewritten to give (5.29).

□□□

The ideal gains K_∞ and G_∞ which minimized the cost (5.29) however do not minimized the performance index J^* in (5.24a) for $q>0$. The performance of the FSWL Kalman filter described in [Williamson (1985)] is also represented using a quadratic cost which is similar to J^* in (5.24a). A first order example also reveals that the optimal FSWL Kalman filter gain is also not the optimal gain which minimizes J^*. In particular, the FSWL-LQG problem is *not* separable. As in [Kwakernaak and Sivan (1972)], the separation principle is guaranteed by the fact that P(1,2)=0 where P(1,2) is the sub-matrix defined by

$$P \triangleq \xi\{\begin{bmatrix}\varepsilon(k)\varepsilon'(k) & \varepsilon(k)\hat{x}'(k)\\ \hat{x}(k)\varepsilon'(k) & \hat{x}(k)\hat{x}'(k)\end{bmatrix}\}$$

$$\triangleq \begin{bmatrix}P(1,1) & P(1,2)\\ P(1,2)' & P(2,2)\end{bmatrix} \qquad (5.33a)$$

From (5.21a-c), we obtain

$$P(1,2) = (\Phi-KL)P(1,2)(\Phi-\Gamma G)' + KLP(1,1)(\Phi-KL)' -$$
$$q(\Phi-KL)(\Phi-\Gamma G)' + qKL(\Phi-KL)' - q\varphi\varphi' - K\Lambda K' \qquad (5.33b)$$

Substitution of K in (5.33b) using (5.6a) for S=P(1,1) will not make P(1,2)=0 in (5.33a) even though both K and G are stabilizing unless when q=0 (ie. infinite

precision arithmetic).

It can be seen in (5.24a) that it is necessary to compute the term $tr(\Upsilon P_2)$ in order to calculate the total cost J^*. In the next section, we shall show the importance of this term when we deal with the structure optimization problem. Equations (5.23c) and (5.24b) permit the computation of $tr(\Upsilon P_2)$. In the following lemma, an alternative method for computing this term is proposed. We shall show in the next section that this lemma is useful for finding the optimal structure [Mullis and Roberts (1976), Hwang (1977), Williamson (1986)].

Lemma 5.3 Consider the term $tr(\Upsilon P_2)$ stated in the cost J^* in (5.24a) with P_2 defined in (5.23c). Consider as well the augmented system matrices Φ_a and Υ described in (5.21b) and (5.24b) respectively. Then, the term $tr(\Upsilon P_2)$ can be rewritten as follows

$$tr(\Upsilon P_2) = tr(I_a' \Psi_a I_a) \tag{5.34a}$$

where

$$I_a' = [I_X \quad -I_X] \tag{5.34b}$$

and where

$$\Psi_a = (\Phi_a - \varphi_a)' W (\Phi_a - \varphi_a) \tag{5.34c}$$

where the square matrix W satisfies the following Lyapunov equation

$$W = \Phi_a' W \Phi_a + \Upsilon \tag{5.34d}$$

Proof: The solution for P_2 defined in (5.23c) is given by

$$P_2 = \sum_{k=0}^{\infty} \Phi_a^k (\Phi_a - \varphi_a) I_a I_a' (\Phi_a - \varphi_a)' (\Phi_a^k)' \tag{5.35a}$$

where I_a in (5.34b) satisfies $I_a = I_a I_a'$. Then using the invariance of the trace operator under transposition (ie. tr(AB)=tr(BA)) we have

$$tr(\Upsilon P_2) = tr(I_a' (\Phi_a - \varphi_a)' \sum_{k=0}^{\infty} (\Phi_a^k)' \Upsilon \Phi_a^k (\Phi_a - \varphi_a) I_a) \tag{5.35b}$$

The solution of the Lyapunov equation (5.34d) is given by

$$W = \sum_{k=0}^{\infty} (\Phi_a^k)' \Upsilon \Phi_a^k \qquad (5.35c)$$

The result stated in (5.34a) can then be obtained from (5.35b–c) and (5.34c).

□□□

Note that the augmented matrix Φ_a is assumed to be stable since we assume that Φ–KL and Φ–ΓG are both stable. Consequently, the solution W defined in (5.34c) exists, is unique and is positive definite provided the pair $\{\Phi_a, \Upsilon_a\}$ is observable where $\Upsilon = \Upsilon_a \Upsilon_a'$.

Consider the transformed FSWL representation which can be derived from (5.11a–c)

$$\hat{z}^*(k+1) = \tilde{\Phi}_2 Q[\hat{z}^*(k)] + \tilde{\varphi}\epsilon(k) + \tilde{K}\tilde{y}_e^*(k) \qquad (5.36a)$$

$$u^*(k) = -Q[\tilde{G}Q[\hat{z}^*(k)]] = -\tilde{G}Q[\hat{z}^*(k)] + \delta(k) \qquad (5.36b)$$

$$Q[\hat{z}^*(k)] = \hat{z}^*(k) - \epsilon(k) \qquad (5.36c)$$

where

$$\tilde{y}_e^*(k) = y(k) - \tilde{L}Q[\hat{z}^*(k)] \qquad (5.36d)$$

$$\tilde{\Phi}_2 = T^{-1}(\Phi - \Gamma G)T \qquad (5.36e)$$

$$\tilde{L} = LT \quad ; \quad \tilde{K} = T^{-1}K \quad ; \quad \tilde{G} = GT \qquad (5.36f)$$

Note that the components of the new residue matrix $\tilde{\varphi}$ are also restricted to be positive power of two. The residues $\epsilon(k)$ and $\delta(k)$ in (5.36b–c) as a result of rounding are not the same as the residues e(k) and d(k) in (5.15b–c) although they will exhibit the same statistical models due to rounding; that is

$$\xi\{\epsilon(k)\} = \xi\{\delta(k)\} = 0$$

$$\xi\{\epsilon_i(k)\delta_j(k)\} = 0 \qquad \text{for all i and J}$$

$$\xi\{\epsilon(k)\epsilon'(k)\} = qI_x$$

$$\xi\{\delta(k)\delta'(k)\} = q$$

The corresponding cost function $\tilde{J}^*$ can be derived by replacing $x^*(k)$ by $z^*(k)$. From lemma 5.1, we obtain

$$\tilde{J}^* = tr(\tilde{\Upsilon}\tilde{P}_1) + q[tr(\tilde{\Upsilon}\tilde{P}_2) + tr(R)] \qquad (5.37a)$$

where

$$\tilde{\Upsilon} = \tilde{T}'\Upsilon\tilde{T} \tag{5.37b}$$

$$\tilde{T} = \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix} \tag{5.37c}$$

and the transformed covariances $\tilde{P}_1$ and $\tilde{P}_2$ are given by

$$\tilde{P}_1 = \tilde{\Phi}_a\tilde{P}_1\tilde{\Phi}_a' + \tilde{K}_a\Theta\tilde{K}_a' + q\tilde{H}\tilde{H}' \tag{5.38a}$$

$$\tilde{P}_2 = \tilde{\Phi}_a\tilde{P}_2\tilde{\Phi}_a' + (\tilde{\Phi}_a-\varphi_a)I_a(\tilde{\Phi}_a'-\varphi_a) \tag{5.38b}$$

where $\tilde{\Phi}_a$, $\tilde{K}_a$ and $\tilde{H}$ can be derived from (5.21b–c) and (5.23e) by using (5.36d–e) and are given by

$$\tilde{\Phi}_a = \tilde{T}^{-1}\tilde{\Phi}_a\tilde{T} \tag{5.38c}$$

$$\tilde{K}_a = \tilde{T}^{-1}\tilde{K}_a \tag{5.38d}$$

$$\tilde{H} = \tilde{T}^{-1}\tilde{H} \tag{5.38e}$$

We mentioned earlier that the compensator performance is affected by both compensator scaling and structure. In the following lemma, we present the relation between the state and the input/output scaling and the compensator performance.

Lemma 5.4 Consider the FSWL compensator (5.19), the performance index (5.24a) and the covariance equations (5.23b–c). Consider as well the transformed FSWL representation (5.36a–e). Then, the effects of scaling on the compensator performance can be stated as follows.

a. The input/output scaling which corresponds to $T\triangleq S=sI_X$ where I_X is an $(n_X x n_X)$ identity matrix for a scalar s alters the cost $\tilde{J}^*$ as follows

$$\tilde{J}^* = tr(\Upsilon P_1) + q\{s^2 tr(\Upsilon P_2) + tr(R)\} \tag{5.39}$$

b. The state vector scaling which corresponds to $T\triangleq S$ where

$$S = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & & \dots & s_{n_X} \end{bmatrix}$$

for scalars s_i for all $i\epsilon[1,n_X]$ changes the index $\tilde{J}^*$ into

$$\tilde{J}^* = tr(\Upsilon P_1) + q\{tr(\tilde{\Upsilon}\tilde{P}_2) + tr(R)\} \tag{5.40a}$$

where $\tilde{\Upsilon}$ and $\tilde{P}_2$ are respectively the scaled version of Υ and P_2 defined in (5.24b) and (5.23c), and are given by

$$\tilde{\Upsilon} = \tilde{S}'\Upsilon\tilde{S} \tag{5.40b}$$

$$\tilde{P}_2 = \tilde{\Phi}_a\tilde{P}_2\tilde{\Phi}_a' + (\tilde{\Phi}_a-\varphi_a)I_a(\tilde{\Phi}_a-\varphi_a)' \tag{5.40c}$$

where

$$\tilde{\Phi}_a = \tilde{S}^{-1}\Phi_a\tilde{S} \tag{5.41a}$$

$$\tilde{S} = \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} \tag{5.41b}$$

and φ_a is given by (5.23d) for $\varphi=I_x$. Furthermore, the scaled version of (5.34a) can be written as

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(S^2 I_a'\Psi_a I_a) \tag{5.42}$$

where Ψ_a and I_a are given by (5.34c) and (5.34b).

Proof: Applying a similarity transformation T=S on the covariance equation P_1 in (5.23b) gives

$$\tilde{P}_1 = \tilde{\Phi}_a\tilde{P}_1\tilde{\Phi}_a' + \tilde{K}_a\Theta\tilde{K}_a' + q\tilde{H}\tilde{H}' \tag{5.43}$$

where $\tilde{\Phi}_a$ and Θ are given by (5.41a) and (5.23e) and $\tilde{K}_a$ and $\tilde{H}$ are given by

$$\tilde{K}_a = \tilde{S}^{-1}K_a \tag{5.44a}$$

$$\tilde{H} = \tilde{S}^{-1}H \tag{5.44b}$$

Substitution of $\tilde{\Phi}_a$, $\tilde{K}_a$ and $\tilde{H}$ in (5.43) using (5.41a) and (5.44a–b), we get

$$\tilde{P}_1 = \tilde{S}^{-1}P_1(\tilde{S}^{-1})' \tag{5.44c}$$

Setting $\tilde{T}=\tilde{S}$ in (5.37b) we obtain (5.40b). From (5.40b) and (5.44c) we have

$$tr(\tilde{\Upsilon}\tilde{P}_1) = tr(\Upsilon P_1) \tag{5.45}$$

a. For of the input/output scaling (ie. $S=sI_x$) from (5.41a–b) we have

$$\tilde{\Phi}_a = \Phi_a \text{ and } \tilde{\varphi}_a = \varphi_a \tag{5.46a}$$

Therefore, from (5.38b) and (5.23c) we have

$$\tilde{P}_2 = P_2 \tag{5.46b}$$

Similarly, substitution of $S=sI_x$ in (5.40b) yields

$$\tilde{\Upsilon} = s^2\Upsilon \tag{5.46c}$$

From (5.37a), (5.45) and (5.46b–c) the cost $\tilde{J}^*$ in (5.39) then follows.

b. For the state vector scaling, from the covariance equation (5.40c), it follows that

$$\tilde{P}_2 \neq \tilde{S}^{-1}P_2(\tilde{S}^{-1})'$$

Therefore, using (5.45) the cost $\tilde{J}^*$ in (5.37a) can be rewritten as in (5.40a). From (5.34c–d) we obtain

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(I_a'\tilde{S}\Psi_a\tilde{S}I_a)$$

but

$$\tilde{S}I_a = I_aS$$

The result in (5.42) then follows.

□□□

The ℓ_2–scaling constraint that we seek to satisfy is to make the variance of each of the compensator states to be equal to the variance of the compensator input. But the compensator input can be scaled independently using the input–output scaling approach as hinted in lemma 5.4. Therefore, the ℓ_2–scaling constraint that we seek to satisfy is

$$\tilde{P}(2,2) = \begin{bmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & & \vdots \\ * & * & \dots & 1 \end{bmatrix} \tag{5.47}$$

where $\tilde{P}(2,2)$ is the scaled version of the sub–matrix P(2,2) defined in (5.33a). In the ideal case (ie. q=0 or B=∞ in (5.17)) the scaling constraint (5.47) for fixed gains K_q and G_q can be achieved by setting

$$S_{jj} = P_{jj}(2,2)^{\frac{1}{2}} \tag{5.48}$$

for all $j=n_x+1, n_x+2, \dots, 2n_x$ where A_{jj} denotes the $(j,j)^{th}$ component of A and P(2,2) is defined in (5.33a). This scaling procedure is in fact the scaling approach which is commonly used in both digital filtering and control [Mullis and Roberts (1976), Hwang (1977), Williamson (1986), Moroney (1983), Moroney et. al (1981)].

The problem of FSWL design for the default structure can be stated as follows.

FSWL Design 1 (ie. Default Structure) Consider the minimal discrete-time model (5.8a-b) and the quadratic cost J^* in (5.9a). Suppose the plant parameters $\{\Phi,\Gamma,L\}$, the noise covariances $\{\Omega,\Lambda\}$, the performance weightings matrices $\{Q,M,R\}$ and the state wordlength B are given. The FSWL design 1 problem can be expressed as the following constrained optimization problem.

$$\min_{\{K,G\}} \; J^* \tag{5.49}$$

where J^* is given by (5.24a), and where K and G are respectively the Kalman filter and the controller gains subject to the unity ℓ_2-scaling constraint (5.47).

□□□

For the FSWL design 1, no algebraic Riccati techniques have been found to aid the solution. This is mainly due to the non-separability of the prediction and control problems mentioned earlier. The optimization has been achieved by a brute force numerical procedure using the augmented Lagrange approach [Gill et. al (1981), Fletcher (1980), Polak (1971)]. However the speed of convergence and probability of obtaining the (possibly) *global* optimum can be improved using the following algorithm.

Algorithm 5.1

1. Initialize the algorithm with $K_0=K_\infty$ and $G_0=G_\infty$ which minimizes $tr(\Upsilon P_1)$ for q=0 (or B=∞) in (5.24a).
2. Let $q=\epsilon$ for a sufficiently 'small' $\epsilon>0$. Find the optimum K_ϵ and G_ϵ by solving the FSWL design 1 for $q=\epsilon$ defined by (5.49).
3. Increase q in small increments and then repeat step 2 until the value corresponding to the desired state wordlength B in the FSWL design 1.

□□□

Note that in the step 2 above even though in the implementation the wordlength B in the FSWL design 1 must be an integer, q can be arbitrary in the constrained

numerical optimization of the cost $\tilde{J}^*$. Heuristically, it is reasonable to expect that the gains K_ϵ and G_ϵ corresponding to the local minimum for $q=\epsilon$ obtained by a gradient procedures beginning with the initial estimates K_0 and G_0 will also provide the global miminum solution for $q=\epsilon$.

The following example which is based on a 6^{th} order model of longitudinal control of a modern transport aeroplane illustrates the FSWL compensator design procedure described above.

Example 5.2 Consider the minimal 6^{th} order model used in example 5.1 where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider also the quadratic cost J^* defined by (5.9a) where the weighting factors Q, M and R are respectively given by (A.23), (A.24) and (A.25).

For the purpose of comparison, we first calculate the performance index J for the ideal gains K_∞ and G_∞ and for different values of the state wordlength B. For each fixed state wordlength B, the compensator input and states are scaled as in lemma 5.4 to achieve the unity ℓ_2–scaling constraint (5.47).

Wordlength B	(K_∞, G_∞)		(K_q, G_q)	
	$tr(\Upsilon P_1)$ (10^{-5})	$qtr(\tilde{\Upsilon}\tilde{P}_2)$ (10^{-5})	$tr(\Upsilon P_1)$ (10^{-5})	$qtr(\tilde{\Upsilon}\tilde{P}_2)$ (10^{-5})
10	9.6025	0.0243	9.6026	0.0092
8	9.6028	0.1518	9.6034	0.0213
6	9.6058	0.3280	9.6131	0.2157
4	9.6544	3.3039	9.7692	1.8236

Table 5.2 The compensator performance resulting from
a. Ideal (q=0) design (denoted by K_∞,G_∞)
b. FSWL design 1 (denoted by K_q,G_q)

In table 5.2, we present the resulting cost J^* in terms of $tr(\Upsilon P_1)$ and $tr(\tilde{\Upsilon}\tilde{P}_2)$. The term tr(R) in the cost J^* is not shown since it is not affected by the choice of

the predictor and the controller gains. For each fixed state wordlength B, the FSWL design 1 solution was sought by means of algorithm 5.1 and of lemmas 5.2, 5.3 and 5.4. The results are presented in table 5.2.

5.4 OPTIMUM FINITE STATE WORDLENGTH STRUCTURE : FIXED GAINS

In the previous section the design of the FSWL compensator for a fixed structure which is called the FSWL Design 1 was examined. This design produces the optimal gains K_q and G_q which minimizes the performance index J^* in (5.40a) subject to the ℓ_2-scaling (5.47) for a given fractional wordlength B (equivalently covariance q defined in (5.17)). In this section, we assume that the gains K_q and G_q for a given state wordlength B and for the default structure with scaling have been found.

Consider the compensator with the transformed structure as defined by (5.36a-f), but now the transformation matrix T is not restricted to be in a special form as to maintain the default structure, even though it has to be selected such that the unity ℓ_2-scaling (5.47) is satisfied. In lemma 5.4, the effects of the scaling matrix S on the quadratic cost J^* were established. For an arbitrary non-singular transformation T, the result in lemma 5.1 and 5.3 can be restated as follows.

Lemma 5.5 Consider the transformed finite state wordlength compensator (5.36a-c) where the transformation matrix T is non singular, and the corresponding cost J^* defined in (5.37a-c) where the covariance matrices $\tilde{P}_1$ and $\tilde{P}_2$ satisfy the Lyapunov equations (5.38a-b). Then, we have the following facts.

a. For an arbitrary non-singular matrix T

$$tr(\tilde{T}\tilde{P}_1) = tr(TP_1) \tag{5.50}$$

where $\tilde{T}$ and $\tilde{P}_1$ are given by (5.37b) and (5.38a), respectively.

b. For an arbitrary unitary matrix T=U (ie.$U=(U^{-1})'$)

$$tr(\tilde{T}\tilde{P}_2) = tr(TP_2) \tag{5.51a}$$

However for an arbitrary non-singular matrix T

$$tr(\tilde{\Upsilon}\tilde{P}_2) \neq tr(\Upsilon P_2) \tag{5.51b}$$

c. For an arbitrary non-singular matrix T

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(T' I_a' \Psi_a I_a T) \tag{5.52}$$

where I_a and Ψ_a are defined by (5.34b-c).

Proof: Subtitution of $\tilde{\Phi}_a$ in (5.38a) using (5.38c) yields

$$\tilde{P}_1 = \tilde{T}^{-1} P_1 (\tilde{T}^{-1})' \tag{5.53a}$$

From (5.37b) and (5.53a), we get (5.50). Substitution of $\tilde{\Phi}_a$ in (5.38c) for T=U and using the fact that

$$\tilde{U} I_a \tilde{U}' = I_a$$

where I_a is defined in (5.23d) and

$$\tilde{U} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$$

the result in (5.51a) then follows. For an arbitrary non-singular matrix T, it can be deduced from (5.38b) that

$$\tilde{P}_2 \neq \tilde{T}^{-1} P_2 (\tilde{T}^{-1})' \tag{5.53b}$$

This proves (5.51b). The transformed version of (5.34a) can be written as follows

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(I_a' \tilde{T}' \Psi_a \tilde{T} I_a) \tag{5.54}$$

but

$$\tilde{T} I_a = I_a T$$

where I_a is defined in (5.34b) and $\tilde{T}$ is defined in (5.37c). This completes the proof.

□□□

It can be seen in (5.37a) that the cost $\tilde{J}^*$ contains three terms. In lemma 5.5, it is shown that the first term of the right hand side of (5.37a) is invariant under the coordinate basis transformation. In fact, this is the property of an ideal compensator (ie. q=0) described in lemma 5.2. The third term in the right hand side of (5.37a) can also be shown to be constant for any choice of non-singular transformation T. In summary, the effects of the compensator structure on the performance can be measured by considering only the changes of the term $tr(\tilde{\Upsilon}\tilde{P}_2)$

defined in (5.52).

FSWL DESIGN 2 (ie. Fixed gains) Consider the minimal discrete-time model (5.8a-b) and the quadratic cost J^* defined in (5.9a). Suppose the FSWL gains K_q and G_q which minimize the quadratic cost J^* for a fixed state wordlength B and a fixed compensator structure are given. The FSWL design 2 problem can be expressed as the following constrained optimization problem.

$$\min_{\{T\}} \text{tr}(\tilde{\Upsilon}\tilde{P}_2) \tag{5.55}$$

subject to the unity ℓ_2-scaling constraint (5.47) where the sub-cost $\text{tr}(\tilde{\Upsilon}\tilde{P}_2)$ is defined by (5.52) and where $\tilde{P}(2,2)$ in (5.47) is the transformed version of $P(2,2)$ defined in (5.33a).

□□□.

The comparison of different structures can be done with respect to the term $\text{tr}(\tilde{\Upsilon}\tilde{P}_2)$ in (5.52) assuming all structures are properly scaled based on the unity ℓ_2-scaling defined in (5.47). From (5.53a-b), we can deduce that an arbitrary non-singular transformation T would not be able to make the sub-matrix $\tilde{P}(2,2)$ to satisfy the unity ℓ_2-scaling constraint (5.47). To overcome this difficulty the covariance matrix P in (5.23a) is approximated by the covariance matrix P_1. This approximation is valid for a large input-signal to quantization-noise ratio (ie. $q \ll 1$ in (5.17)) in which case the term qP_2 in (5.23a) can be neglected. In fact, this is the approach commonly used either in the signal processing context [Mullis and Roberts (1976), Williamson (1986)] or in the control context [Moroney et. al (1980), Scharf and Sigudson (1984), Sasahara et. al (1984), Ahmed and Belanger (1984a)]. After a desired transformation T has been found, the original unity ℓ_2-scaling constraint (5.47) can be satisfied by using a certain unitary transformation U described in lemma 5.5. Note that a unitary transformation U satisfies both (5.50) and (5.51a). We will discuss these scaling procedures later in this section. With this approximation, the scaling constraint that we seek to satisfy is given by

$$\tilde{P}_1(2,2) = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \cdots & * \\ \vdots & \vdots & & \vdots \\ * & * & \cdots & 1 \end{bmatrix} \tag{5.56a}$$

where $\tilde{P}_1(2,2)$ is $(n_x x n_x)$ (where n_x is the dimension of the state x(k)), and is given by

$$\tilde{P}_1 = \left[\begin{array}{c|c} \tilde{P}_1(1,1) & \tilde{P}_1(1,2) \\ \hline \tilde{P}_1(2,1) & \tilde{P}_1(2,2) \end{array}\right] \tag{5.56b}$$

In the digital filter context, this structure optimation problem has been solved both for zero and non-zero integer residue matrix φ. For the case of no IRC (ie. φ=0) the optimal solution of the structure optimization problem can be found in [Mullis and Robert (1976), Hwang (1977)] while the solution for a non-zero φ is presented in [Williamson (1986)]. Since we do not intend to restrict the integer residue matrix φ to be zero, the latter approach (which fixes $\varphi=I_x$) is more suitable in this context.

From (5.56b), for an arbitrary non-singular matrix T we obtain

$$\tilde{P}_1(2,2) = T^{-1}P_1(2,2)(T^{-1})' \tag{5.57}$$

From (5.52) and (5.57), the following result can be derived.

Lemma 5.6 Consider the sub-cost $tr(\tilde{\Upsilon}\tilde{P}_2)$ defined in (5.52) and the covariance matrix $\tilde{P}_1(2,2)$ defined in (5.57). Assume the residue matrices φ and $\tilde{\varphi}$ defined in (5.19) and (5.36a) are both identity matrices. Then, the eigenvalues $\{\rho_k^{-2}\}$ of $P_1(2,2)I_a'\Psi_a I_a$ are invariant under coordinate basis transformation.

Proof: From (5.34c) and (5.57) we have

$$\tilde{P}_1(2,2)I_a'\tilde{\Psi}_a I_a = T^{-1}P_1(2,2)I_a'\Psi_a I_a T \tag{5.58}$$

The gains K_q and G_q produced by the FSWL design 1 are stabilizing, and assuming Φ_a defined in (5.21b) has no modes on the unit circle, we conclude $\Psi_a>0$ in (5.34c) and $P_1(2,2)>0$ in (5.23b). To show $\Psi_a>0$, choose T such that W=I in

(5.34c). Therefore eigenvalues $\bar{\rho_k}^2 > 0$ for all $k=1,2,\ldots,n_x$.

□□□

The effects of applying an arbitrary transformation T (T^{-1} exists) to a given structure $\{\Phi,\Gamma,L,K,G\}$ on the compensator performance can be analyzed by considering the singular value decomposition (SVD) [Bellman (1970)] of the non-singular matrix T; that is

$$T = R_1 \Pi R_0' \tag{5.59a}$$

where R_1 and R_0 are both unitary matrices and where

$$\Pi = \begin{bmatrix} \pi_1 & 0 & \ldots & 0 \\ 0 & \pi_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \pi_{n_x} \end{bmatrix} \tag{5.59b}$$

where $\pi_i > 0$ for all $i \epsilon [1,n_x]$. Substitution of T in (5.52) using (5.59a) yields

$$\begin{aligned} tr(\tilde{T}\tilde{P}_2) &= tr(\Pi^2 R_1' I_a' \Psi_a I_a R_1) \\ &= \sum_{i=1}^{n_x} \pi_i d_i \end{aligned} \tag{5.60a}$$

where π_i for all $i \epsilon [1,n_x]$ are given in (5.59b) and where

$$R_1' I_a \Psi_a I_a R_1 = \begin{bmatrix} d_1 & * & \ldots & * \\ * & d_2 & \ldots & * \\ \vdots & \vdots & & \vdots \\ * & * & \ldots & d_{n_x} \end{bmatrix} \tag{5.60b}$$

Substitution of T in (5.57) using (5.59a) gives

$$\tilde{P}_1(2,2) = R_0 \Pi^{-1} R_1' P_1(2,2) R_1 \Pi^{-1} R_0' \tag{5.61}$$

The following two lemmas which are useful in deriving an optimal transformation T are adopted from [Mullis and Roberts (1976), Hwang (1977)].

Lemma 5.7 [Mullis and Roberts (1976), Hwang (1977)]. Given an $(n_x x n_x)$ diagonal positive-definite matrix E. There exists a (non-unique) orthogonal matrix R_o such that

$$R_o E R_o = \begin{bmatrix} e & * & \dots & * \\ * & e & \dots & * \\ \vdots & \vdots & & \vdots \\ * & * & \dots & e \end{bmatrix} \tag{5.62a}$$

if and only if; for a (positive) scalar e

$$tr(E) = e n_x \tag{5.62b}$$

□□□

Lemma 5.8 [Mullis and Roberts (1976), Hwang (1977)]. Given unitary matrices R_o and R_1 and an $(n_x x n_x)$ diagonal matrix Π where its diagonal elements are given by $\{\pi_i; 0 \le i \le n_x\}$. Consider the eigenvalues $\{\rho_k^{-2}\}$ of the positive-definite matrix $P_1(2,2) I_a' \Psi_a I_a$ defined in lemma 5.6. Then, subject to the unity ℓ_2-scaling constraint

$$R_o \Pi^{-2} R_o' = \begin{bmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & & \ddots & \vdots \\ * & * & \dots & 1 \end{bmatrix} \tag{5.63a}$$

the lower bound of (5.60a) is given by

$$tr(\Pi^2 R_1' I_a' \Psi_a I_a R_1) \ge \frac{(\sum_{i=1}^{n_x} \rho_i)^2}{n_x} \tag{5.63b}$$

where the minimum value denoted by the equality in (5.63b) is achieved if only if (5.60b) is diagonal.

□□□

Given $P_1(2,2)$ and $I_a'\Psi_a I_a$ where Ψ_a is defined in (5.34c), the covariance matrix $P_1(2,2)$ can be transformed into an identity matrix by a certain equivalent transformation T_0 [Bellman (1970)]; that is

$$\tilde{P}_1(2,2) \triangleq T_0^{-1}P_1(2,2)(T_0^{-1})' = I \tag{5.64a}$$

Consequently, the sub-cost $I_a'\Psi_a I_a$ is transformed into

$$I_a'\Psi_a I_a \triangleq T_0' I_a'\Psi_a I_a T_0 \tag{5.64b}$$

Note that the transformation (5.64a) implies

$$T_0 T_0' = P_1(2,2) \tag{5.65}$$

This particular realization obviously satisfies the unity ℓ_2-scaling constraint (5.56a). This structure has been used in [Mullis and Roberts (1976), Hwang (1977), Williamson (1986)] as a convenient initial structure in seeking the optimal (ie. minimum round-off noise) structure.

Theorem 5.1 [Williamson (1986)] Consider the minimal discrete-time model (5.8a-b) and the corresponding transformed FSWL compensator (5.36a-c) and the cost function $\tilde{J}^*$ defined in (5.37a). Consider as well the unity ℓ_2-scaling (5.56a) and an identity residue matrix φ. Then, the equivalent transformation T which achieves the minimum in (5.55) subject to the unity ℓ_2-scaling (5.56a) is given by

$$T = T_0 R_1 \Pi R_0' \tag{5.66}$$

where T_0 is given by (5.64a-b), Π is a diagonal matrix of the form

$$\Pi = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \pi_{n_x} \end{bmatrix} \tag{5.67a}$$

where n_x is the dimension of the state $x^*(k)$, and where

$$\pi_i = \left[\frac{\sum_{j=1}^{n_x} \rho_j}{n_x \rho_i} \right]^{\frac{1}{2}} \tag{5.67b}$$

and the unitary matrices R_0 and R_1 satisfy

$$R_0 \Pi^{-2} R_0' = \begin{bmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & & \vdots \\ * & * & \dots & 1 \end{bmatrix} \tag{5.67c}$$

$$R_1' T_0' I_a' \Psi_a I_a T_0 R_1 = \begin{bmatrix} \rho_1^2 & 0 & \dots & 0 \\ 0 & \rho_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \rho_{n_x}^2 \end{bmatrix} \tag{5.67d}$$

Moreover, the minimum sub-cost defined in (5.54) is given by

$$tr(\tilde{\Upsilon}\tilde{P}_2) = \frac{(\sum_{i=1}^{n_x} \rho_i)^2}{n_x} \tag{5.67e}$$

Proof: The result stated in (5.67c-d) and (5.67e) follows from lemma 5.7 for e=1 and lemma 5.8.

□□□

As we mentioned earlier, the similarity transformation T described in theorem 5.1 does not fully satisfy the unity ℓ_2-scaling constraint (5.47) since the derivation was based on the approximation of the covariance matrix P defined in (5.23a). The scaling constraint (5.47) can be satisfied by using an iterative procedure which uses the unitary matrix defined in lemma 5.7 and an input/output scaling (ie. $S=sI_x$ described in lemma 5.4).

Algorithm 5.2

1. Given a fixed state wordlength B (and hence q in (5.17)) and the gains K_q and G_q. Find the equivalent transformation T which achieves the minimum in (5.55) subject to the unity ℓ_2-scaling (5.56a) by means of

theorem 5.1.

2. Find a unitary matrix U (using an iterative procedure developed in [Mullis and Roberts (1976)]) such that the diagonal elements of the covariance matrix $\tilde{P}(2,2)$ are all equal to the factor e where $e=tr(\tilde{P}(2,2)/n_x)$.
3. Apply an input/output scaling $S=e^{\frac{1}{2}}I_x$ on the covariance equations (5.23b–c).
4. Repeat step 2 and 3 until the unity ℓ_2–scaling (5.47) is satisfied.

□□□

Example 5.3. Consider the minimal 6^{th} order model used in examples 5.1–2 where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider also the quadratic cost J^* defined by (5.22a–b) where the weighting matrices Q, M and R are respectively given by (A.23), (A.24) and (A.25).

Using the ideal (q=0) gains K_∞ and G_∞ given in (A.28) and (A.29), for each state wordlength B we calculated the term $tr(\tilde{T}\tilde{P}_2)$ defined in (5.52) for the default structure subject to the unity ℓ_2–scaling constraint (5.47). This can be done by means of the algorithm 5.2 for $T=I_x$ in step 1 of the algorithm. Then, we carried out the structure optimization by means of the algorithm 5.2 for each state wordlength B and for the ideal gains K_∞ and G_∞. The following matrix is the optimal transformation T^0 which achieves the minimum in (5.55) for a state wordlength B=4.

T^0 =

```
  COLUMNS    1 THRU    3
-6.134541332883778D+00  -4.644366181598011D-02   3.174135154095813D+00
 3.242067095423542D-01   6.748691532218376D-02  -4.027431195052797D-01
 2.200511240320468D-02   1.055905694117085D-02   2.085462602965967D-02
-1.060748181099880D+00   8.173062418683313D-02  -8.054957539969604D-01
 2.400729814682667D-02  -4.901143134848550D-01   1.614325626127295D-01
 2.916645132799563D-01  -2.920286311359723D-01   1.349449577936815D-01

  COLUMNS    4 THRU    6
-1.067800804151486D+01   4.562357980677966D+00   9.475156349531173D+00
 2.423313931368735D-02   1.663034580469126D-01  -1.440642465763055D-01
 2.129811582134363D-02   2.079901835151734D-02   2.141288682038899D-02
 7.995853555043635D-01  -1.720587428849018D-01   1.147201876453720D+00
 1.089778070001790D-01   8.589611731548165D-02   9.045096384048130D-02
-7.858557287222517D-02  -3.928090401972161D-01   2.789763103125468D-01
```

The term $tr(\tilde{T}\tilde{P}_2)$ for each state wordlength B and for both the default and the optimum structures are presented in table 5.3.

Wordlength B	$qtr(\tilde{T}\tilde{P}_2)$	
	Default Structure (10^{-6})	Algorithm 5.2 (10^{-6})
10	0.2431	0.2411
8	1.5183	1.3989
6	3.2796	2.6886
4	33.0387	24.9771

Table 5.3 The compensator performance resulting from
a. Ideal (q=0) design : default structure
b. FSWL design 2 : fixed gains (Algorithm 5.2)

5.5 OPTIMUM FINITE STATE WORDLENGTH LQG REGULATOR

In the last two sections, the FSWL–LQG regulator problem was examined but in each section only a part of the problem was solved. First, the compensator structure is fixed at the default structure, the design which is called the FSWL design 1 is aimed at finding the Kalman filter and the controller gains K_q and G_q for a certain state wordlength B which minimizes the performance index J^* subject to unity ℓ_2–scaling (5.47). The objective of the FSWL design 2 is to find an equivalent (non–singular) transformation T such that the performance index J^* represented by the sub–cost $tr(\tilde{T}\tilde{P}_2)$ is mimimized subject to unity ℓ_2–scaling constraint (5.56a) for fixed gains K_q and G_q.

The complete FSWL–LQG design requires simultaneous optimization of both the compensator gains (ie. K_q and G_q) and the compensator structure (ie. T). This can be achieved in an iterative manner as follows.

Algorithm 5.3

1. With the given plant model (5.8a–b) and the state wordlength B (and hence q in (5.17)).
2. Optimize the gains K_q and G_q (ie. FSWL design 1) by means of algorithm 5.1 subject to the unity ℓ_2–scaling (5.47). This results in
   a. the scaled system matrices $\{\Phi_s,\Gamma_s,L_s\}$ (5.68a)
   b. the scaled noise covariances $\{\Omega_s,\Lambda\}$ (5.68b)
   c. the scaled weighting matrices $\{Q_s,M_s,R\}$ (5.68c)
3. With the gains K_q and G_q from the previous step, and with the scaled system matrices, the scaled noise covariance and the scaled weighting matrices defined in (5.68a–c), optimize the equivalent transformation T (ie. FSWL design 2) using algorithm 5.2. This results in
   a. the transformed system matrices $\{\Phi_t,\Gamma_t,L_t\}$ (5.68d)
   b. the transformed covariances $\{\Omega_t,\Lambda\}$ (5.68e)
   c. the transformed weighting matrices $\{Q_t,M_t,R\}$ (5.68f)
   d. the transformed gains $\{K_t,G_t\}$ (5.68g)
4. With the transformed system matrices, the transformed covariances and the transformed weighting matrices defined in (5.68d–f), go to step 2.

□□□

Example 5.4 Consider again the minimal 6^{th} order model used in examples 5.1–3 where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider also the quadratic cost J^* defined by (5.22a–b) where the weighting matrices Q, M and R are respectively given by (A.23), (A.24) and (A.25).

For each state wordlength B (and hence q in (5.17)), the optimal FSWL compensator (ie. the optimal $\{K_q,G_q,T\}$) was designed using the algorithm 5.3. The resulting performance index J^* for each state wordlength B is presented in table 5.4. As a comparison, we also present the quadratic cost J^* computed in example 5.2

for the default structure with scaling and for the ideal (ie. q=0 in (5.17)) gains K_∞ and G_∞ given in (5.28) and (5.29).

wordlength B	$J^*(10^{-5})$ (ideal)	$J^*(10^{-5})$ (algorithm 5.3)
12	9.6030	9.6029
10	9.6276	9.6160
8	9.7667	9.3124
6	10.1372	9.7236
4	16.2136	13.5386

Table 5.4 The compensator performance resulting from
a. Ideal (q=0) design
b. FSWL-LQG design (ie. algorithm 5.3)

5.6 COEFFICIENT WORDLENGTH CONSIDERATION IN FINITE STATE WORDLENGTH LINEAR QUADRATIC GAUSSIAN REGULATOR

In sections 5.3, 5.4 and 5.5, it was assumed that the compensator coefficients are implemented using a large (but finite) wordlength B_c. In this section, we consider the effects of the *finite coefficient wordlength* (FCWL) on the performance of the compensators designed in the earlier sections. First, assume that the compensator states are implemented using an arbitrary large (but finite) wordlength B. This implies that the degradation in the compensator performance is only due to the coefficient quantization. The FCWL compensator can be represented (and hence implemented) in several ways. Three examples listed below are the FCWL version of the FWL compensators (5.10a-c).

$$\hat{x}^*(k+1) = (\Phi-\Gamma G-KL)^*\hat{x}^*(k) + K^*y^*(k) \qquad (5.69a)$$

$$\hat{x}^*(k+1) = (\Phi-KL)^*\hat{x}^*(k) + \Gamma^*u^*(k) + K^*y^*(k) \qquad (5.69b)$$

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)^*\hat{x}^*(k) + K^*(y^*(k)-L^*\hat{x}^*(k)) \qquad (5.69c)$$

where the coefficient A^* means the quantized version of the ideal coefficient A to a

B_c-bit fractional representation, and

$$u^*(k) = - G^*\hat{x}^*(k) \tag{5.70}$$

and where

$$\Phi^*_{c\ell} \triangleq (\Phi-\Gamma G-KL)^* = \Phi_{c\ell} + \Delta\Phi_{c\ell} \tag{5.71a}$$

$$\Phi^*_1 \triangleq (\Phi-KL)^* = \Phi_1 + \Delta\Phi_1 \tag{5.71b}$$

$$\Phi^*_2 \triangleq (\Phi-\Gamma G)^* = \Phi_2 + \Delta\Phi_2 \tag{5.71c}$$

$$K^* = K + \Delta K \quad ; \quad \Gamma^* = \Gamma + \Delta\Gamma \tag{5.71d}$$

$$L^* = L + \Delta L \quad ; \quad G^* = G + \Delta G \tag{5.71e}$$

where $\Delta\Phi_{c\ell}$, $\Delta\Phi_1$, $\Delta\Phi_2$, ΔK, $\Delta\Gamma$ and ΔL are residue matrices due to the coefficient quantization which are determined by the coefficient wordlength B_c and the coefficient matrices $\Phi_{c\ell}$, Φ_1, Φ_2, K, Γ and L.

For an infinite (state and) coefficient wordlength, the performance of compensator realized using (5.69a) is the same with the performance of realization (5.69b) or (5.69c). The effects of realizations (5.69a–c) on the compensator performance will be illustrated later by means of an example. In this section, the coefficient quantization analysis will be based on representation (5.69c). The quantization analysis for other realizations can be done in a similar fashion. The corresponding plant representation is given by (5.8a–b), and the performance index is defined in (5.9a).

Note that double precision additions (of fractional width $B+B_c$) is implied in (5.69) therefore $\hat{x}^*(k+1)$ has a $B+B_c$ bit fractional representation. Even though only the quantized $\hat{x}^*(k+1)$ is stored for subsequent calculations, the quantization residue can be neglected since the fractional wordlength B is assumed (arbitrarily) large.

The coefficient quantization analysis which is studied via the residue matrices (5.71a–e) can be done using a deterministic or a statistic approach [Kawamata and Higuchi (1985)]. In this section we wish to examine the finite coefficient wordlength effects on the FSWL compensator performance the statistical approach. Therefore, the statistical analysis is used even though the result would not be as precise as that of the deterministic analysis [Kawamata and Higuchi (1985), Sasahara et. al (1984)]

since in the statistical analysis, the quantization residues are the approximation of the real residues (which are deterministic). Another advantage of using the statistical analysis is that the structure that minimizes the degradation of compensator performance due to state quantization also minimizes the degradation due to coefficient quantization [Sasahara et. al (1984)]. The same conclusion has also been drawn in digital filtering [Fettweis (1973), Jackson (1976), Jackson et. al (1979), Antonio et. al (1983)]. Note that this result is only true based on an approximate analysis.

In the statistical approach, assuming that the process and measurement disturbances are 'white', the coefficient quantization residues can approximately be modelled as zero-mean independent random variables [Knowles and Olcayto (1968), Avenhaus (1972), Crochiere (1975)]. In the following analysis, we assume that the coefficient quantization residues (5.71a-e) are statistically independent and uniformly distributed in the range

$$[(-1/2)2^{-B_c}, (1/2)2^{-B_c}]$$

where B_c is a coefficient wordlength. A justification of this statistical approach for modelling coefficient errors can be found in [Kawamata and Higuchi (1985)].

Define the state prediction error

$$\varepsilon^*(k) \triangleq x^*(k) - \hat{x}^*(k) \tag{5.72}$$

Then, from (5.8a-b) and (5.69c) we get the prediction error equation which is described by

$$\varepsilon^*(k+1) = (\Phi - K^*L)\varepsilon^*(k) - (\Delta\Phi_2 + \Gamma\Delta G + K^*\Delta L)\hat{x}_k^*(k) + \omega(k) - K^*\eta(k) \tag{5.73}$$

where $\Delta\Phi_2$, ΔL and ΔG are defined by (5.71c) and (5.71e) respectively. From (5.69c) and (5.73), we obtain

$$\begin{bmatrix} \varepsilon^*(k+1) \\ \hat{x}^*(k+1) \end{bmatrix} = \Phi_a^* \begin{bmatrix} \varepsilon^*(k) \\ \hat{x}^*(k) \end{bmatrix} + K_a^* \begin{bmatrix} \omega(k) \\ \eta(k) \end{bmatrix} \tag{5.74a}$$

where

$$\Phi_a^* = \begin{bmatrix} \Phi - K^*L & -(\Delta\Phi_2 + \Gamma\Delta G + K^*\Delta L) \\ K^*L & \Phi_2^* - K^*\Delta L \end{bmatrix} \tag{5.74b}$$

$$K_a^* = \begin{bmatrix} I_x & -K^* \\ 0 & K^* \end{bmatrix} \tag{5.74c}$$

where I_x is an $(n_x x n_x)$ identity matrix. Define

$$\Delta K_a \triangleq K_a^* - K_a \tag{5.75a}$$

$$\Delta\Phi_a \triangleq \Phi_a^* - \Phi_a \tag{5.75b}$$

where Φ_a and K_a are given by (5.21b–c). From (5.74b–c) and (5.21b–c), we obtain

$$\Delta K_a = \begin{bmatrix} 0 & -\Delta K \\ 0 & \Delta K \end{bmatrix} \tag{5.75c}$$

$$\Delta\Phi_a = \begin{bmatrix} -\Delta KL & -(\Delta\Phi_2 + \Gamma\Delta G + K^*\Delta L) \\ \Delta KL & \Delta\Phi_2 - K^*\Delta L \end{bmatrix} \tag{5.75d}$$

Define

$$\Delta\epsilon(k) \triangleq \epsilon^*(k) - \epsilon(k) \tag{5.76a}$$

$$\Delta\hat{x}(k) \triangleq \hat{x}^*(k) - \hat{x}(k) \tag{5.76b}$$

where $\epsilon(k)$ and $\hat{x}(k)$ are respectively the ideal prediction error and the ideal compensator state defined in (5.5a) and (5.4a). Subtracting the ideal augmented representation (5.5b) from (5.74a) and after simplification, we obtain

$$\begin{bmatrix} \Delta\epsilon(k+1) \\ \Delta\hat{x}(k+1) \end{bmatrix} = \Phi_a \begin{bmatrix} \Delta\epsilon(k) \\ \Delta\hat{x}(k) \end{bmatrix} + \Delta\Phi_a \begin{bmatrix} \epsilon^*(k) \\ \hat{x}^*(k) \end{bmatrix} + \Delta K_a \begin{bmatrix} \omega(k) \\ \eta(k) \end{bmatrix} \tag{5.77}$$

where ΔK_a and $\Delta\Phi_a$ are given by (5.75c–d).

Substitution of $x^*(k)$ and $u^*(k)$ in the quadratic cost (5.9a) using (5.72) and (5.70), after some rearangement we get the new representation of the index J^* as in (5.22a) where the function s(k) is given by

$$s(k) = \varepsilon^*(k)'Q\varepsilon^*(k) + 2\varepsilon^*(k)'(Q-MG^*)\hat{x}^*(k) + (\hat{x}^*(k)'(Q-2MG^*+G^*R(G^*)')\hat{x}^*(k) \quad (5.78a)$$

where G^* is defined in (5.71e). Furthermore, the function s(k) can be rewritten as

$$s(k) = \left\{ \begin{bmatrix} \varepsilon^*(k)' \\ \hat{x}^*(k)' \end{bmatrix} \Upsilon^* [\varepsilon^*(k) \ \hat{x}^*(k)] \right\} \quad (5.78b)$$

where

$$\Upsilon^* = \begin{bmatrix} Q & Q-MG^* \\ (Q-MG^*)' & Q-2MG^*+(G^*)'RG^* \end{bmatrix} \quad (5.78c)$$

Define

$$\Delta\Upsilon \triangleq \Upsilon^* - \Upsilon \quad (5.79a)$$

Then, by subtracting (5.24b) from (5.78c) we obtain

$$\Delta\Upsilon = \begin{bmatrix} 0 & -M\Delta G \\ -(M\Delta G)' & -2M\Delta G+\Delta G'RG+G'R\Delta G+\Delta G'R\Delta G \end{bmatrix} \quad (5.79b)$$

Lemma 5.9 [Kawamata and Higuchi (1985)] Let ΔZ be an $(n_x x n_x)$ residue matrix defined by

$$\Delta Z \triangleq Z - Z^*$$

where the $(n_x x n_x)$ matrices z and z^* are respectively the infinite precision and the quantized coefficient matrices. Then, for a given square matrix A and a coefficient wordlength B_c

$$\xi\{\Delta Z A \Delta Z'\} = q \begin{bmatrix} \sum_{i=1}^{n_x} r(Z_{1i})a_{ii} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n_x} r(Z_{2i})a_{ii} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \cdots & & \sum_{i=1}^{n_x} r(Z_{n_x i})a_{ii} \end{bmatrix} \quad (5.80a)$$

where Z_{ij} and a_{ij} denote the $(i,j)^{th}$ element of the matrices Z and A and

$$q = \frac{1}{12} 2^{-2B_c} \tag{5.80b}$$

and where the function $r(Z_{ij})$ for all $i\epsilon[1,n_x]$ and $j\epsilon[1,n_x]$ is given by

$$r(Z_{ij}) = \begin{cases} 1 & \text{if } Z_{ij} \text{ is non-integer} \\ 0 & \text{if } Z_{ij} \text{ is integer} \end{cases} \tag{5.80c}$$

Proof: In [Knowles and Olcayto (1968)], the scalar ΔZ_{ij} is 'approximately' modelled as an independent zero-mean random variable with variance

$$\xi\{\Delta Z_{ij}^2\} = qr(Z_{ij}) \tag{5.81a}$$

where q and $r(Z_{ij})$ are given in (5.80b-c). The (assumed) independence of residue elements ΔZ_{ij} for all $i\epsilon[1,n_x]$ and $j\epsilon[1,n_x]$ implies

$$\xi\{\Delta Z_{ij}\Delta Z_{mn}\} = 0 \tag{5.81b}$$

for $i\neq m$ and $j\neq n$, carry out the multiplication in $\{.\}$ on the left hand side of (5.80a) and make use the relations (5.81a-b) to give the result.

□□□

In section 5.2, the ideal cost J was derived. The performance index of the compensator realized using a finite coefficient wordlength is given by J^* in (5.22a) where the corresponding function s(k) is defined in (5.78a). From these two quadratic costs, the effects of finite coefficient wordlength on the compensator performance can be examined. In the following lemma, we present a formulae for measuring the effects of finite coefficient wordlength on the compensator performance. Recall that the models are assumed to be SISO. The method which is first developed in digital filtering [Kawamata and Higuchi (1985)] is similar to the one used in [Sasahara et. al (1984)].

Lemma 5.10 Consider the infinite precision quadratic cost J defined in (5.3) and the FCWL compensator performance index J^* defined in (5.22a) with the function s(k) in (5.78a-b). Consider also the residue matrices defined in (5.71a-e), (5.75a-d) and (5.79a-b). Suppose $\Delta\hat{J}$ is a measure of degradation of the compensator performance due to a finite coefficient wordlength implementation. Then, $\Delta\hat{J}$ is given by

$$\Delta\hat{J} = qR \sum_{j=1}^{n_x} L_G(j,j) + qtr(WZ) \tag{5.82a}$$

where q and R are defined in (5.80b) and (5.3) and where

$$W = \Phi_a' W \Phi_a + \Upsilon \tag{5.82b}$$

$$L_G = \begin{bmatrix} r(G(1)) & 0 & \cdots & 0 \\ 0 & r(G(2)) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r(G(n_x)) \end{bmatrix} \tag{5.82c}$$

where G is the controller gain given in (5.71e), Φ_a is defined in (5.21b) and Υ is defined in (5.24b),

$$Z = (\sigma_y^2 L_K + L_\Phi) I_a + L_g \Gamma \Gamma' \begin{bmatrix} I_x & 0 \\ 0 & 0 \end{bmatrix} \tag{5.82d}$$

where I_a is defined in (5.23d), C is given in (5.71d) and

$$Lg = \sum_{j=1}^{n_x} r(G(j)) \tag{5.82e}$$

$$L_K = \begin{bmatrix} r(K(1)) & 0 & \cdots & 0 \\ 0 & r(K(2)) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r(K(n_x)) \end{bmatrix} \tag{5.82f}$$

where K is the Kalman filter gain given in (5.71d) and

$$L_\Phi = \begin{bmatrix} \sum_{j=1}^{n_x} r(\Phi_2(1,j)) & 0 & \dots & 0 \\ 0 & \sum_{j=1}^{n_x} r(\Phi_2(2,j)) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sum_{j=1}^{n_x} r(\Phi_2(n_x,j)) \end{bmatrix} \quad (5.82g)$$

where Φ_2 is defined in (5.71c) and X(i) for all $i\epsilon[1,n_x]$ are the i^{th}–element of the vector X, and Y(i,j) for all $i\epsilon[1,nx]$ and $j\epsilon[1,n_x]$ are the $(i,j)^{th}$–element of the matrix Y and where

$$\sigma_y^2 \triangleq \xi\{y^*(k)^2\} \quad (5.82h)$$

and the function r(.) is defined by (5.80c).

Proof: Subtract J in (5.3) from J^* in (5.22a) to give

$$\Delta\hat{J} = \Delta J_1 + \Delta J_2 \quad (5.83a)$$

where

$$\Delta J_1 = tr(\Upsilon\xi\{\begin{bmatrix} \Delta\epsilon(k) \\ \Delta\hat{x}(k) \end{bmatrix}[\Delta\epsilon'(k) \quad \Delta\hat{x}'(k)]\}) \quad . \quad (5.83b)$$

$$\Delta J_2 = tr(\xi\{\Delta\Upsilon\}\xi\{\begin{bmatrix} \epsilon^*(k) \\ \hat{x}^*(k) \end{bmatrix}[\epsilon^*(k)' \quad \hat{x}^*(k)']\}) \quad (5.83c)$$

Note that to get (5.83b–c) the independence of the residues $\Delta\epsilon(k)$ and $\Delta\hat{x}(k)$ has been used. Substitute $\Delta\Upsilon$ in (5.83c) using (5.79b) and evaluate the mathematical expectation, we obtain after simplification

$$\Delta J_2 = qRtr(\xi\{\hat{x}^*(k)\hat{x}^*(k)'\}L_G) \quad (5.84a)$$

where L_G is given by (5.82c) and it was derived by means of lemma 5.9. The unity ℓ_2–scaling (5.47) implies that all diagonal elements of $\xi\{.\}$ in (5.84a) are unity. Therefore, ΔJ_2 in (5.84a) can be simplified as

$$\Delta J_2 = qR \sum_{j=1}^{n_x} L_G(j,j) \tag{5.84b}$$

This proves the first term of ΔJ in (5.82a). Now define

$$z_a(k) \triangleq \Delta K \begin{bmatrix} y^*(k) \\ \hat{x}^*(k) \end{bmatrix} \tag{5.85a}$$

where

$$\Delta K = \begin{bmatrix} -\Delta K & -\Delta\Phi_2 - \Gamma\Delta G - K\Delta L + \Delta K L \\ \Delta K & \Delta\Phi_2 + K\Delta L - \Delta K L \end{bmatrix} \tag{5.85b}$$

From (5.77), in steady-state we have

$$\xi\left\{\begin{bmatrix} \Delta\epsilon(k) \\ \Delta\hat{x}(k) \end{bmatrix} [\Delta\epsilon'(k) \;\; \Delta\hat{x}'(k)]\right\} = \sum_{k=-\infty}^{\infty} \Phi_a^k Z (\Phi_a')^k \tag{5.85c}$$

where

$$Z = \xi\{z_a(k) z_a'(k)\} \tag{5.85d}$$

where $z_a(k)$ is defined in (5.85a). Substitution of (5.85c) into (5.83b) and after some rearrangement yields

$$\Delta J_1 = tr(WZ) \tag{5.85e}$$

Substitution of ΔJ_1 and ΔJ_2 in (5.83a) using (5.85e) and (5.83d) produces $\Delta\hat{J}$ in (5.82a). The following relations which are useful for simplifying the representation of Z in (5.85d) are derived by means of lemma 5.9 and the unity ℓ_2-scaling (5.47)

$$\xi\{\Delta K y^*(k) y^*(k) \Delta K'\} = q\sigma_y^2 L_K \tag{5.86a}$$

$$\xi\{\Delta\Phi_{c\ell}\hat{x}^*(k)\hat{x}^*(k)'\Delta\Phi_{c\ell}'\} = qL_\Phi \tag{5.86b}$$

$$\xi\{\Gamma\Delta G\hat{x}^*(k)\hat{x}^*(k)'\Delta G'\Gamma'\} = qL_G\Gamma\Gamma' \tag{5.86c}$$

where L_G, L_K and L_Φ are respectively defined by (5.82c) and (5.82f-g). Carry out the multiplications inside $\{.\}$ in (5.85d), use the independence of the coefficient quantization residues (5.81b) and make use the relations (5.86a-c) to give the simplified Z given in (5.82d). This completes the proof.

□□□

As we mentioned earlier, different representations of a digital compensator (see (5.69a–c)) give a different degradation of the performance. In the following example, we show the effects of a finite coefficient wordlength implementation on the compensator performance.

Example 5.5 Consider the minimal 6^{th} order model used in examples 5.1–4 where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider also the quadratic cost J^* defined by (5.9a) where the weighting factors Q, M and R are respectively given by (A.23), (A.24) and (A.25).

B_c	Φ-KL-ΓG		Φ-KL		Φ-ΓG	
	ΔJ (10^{-7})	$\Delta\hat{J}$ (10^{-7})	ΔJ (10^{-7})	$\Delta\hat{J}$ (10^{-7})	ΔJ (10^{-7})	$\Delta\hat{J}$ (10^{-7})
18	0.0002	0.0002	0.0002	0.0006	0.0000	0.0001
16	0.0236	0.0841	0.0431	0.1012	0.0168	0.0642
14	1.1246	2.4136	1.2231	3.9672	1.0367	2.2736
12	u	u	u	u	u	u

Table 5.5 The degradation of the compensator performance due to a finite coefficient wordlength, for three different compensator described in (5.69a–c).

For each coefficient wordlength B_c, the degradation $\Delta\hat{J}$ was computed by means of lemma 5.10 for the compensator realized using the default structure with scaling and for the ideal gains K_∞ and G_∞ given in (A.28) and (A.29). Also for each coefficient wordlength B_c, the actual degradation ΔJ defined in (5.9b) was calculated; this was done by subtracting the ideal cost J from the (quantized coefficient) cost J^* defined in (5.22a–b). In calculating the degradation ΔJ and $\Delta\hat{J}$ for each realization (5.69a), (5.69b) and (5.69c) we had to employ different scaling approaches. For

realization (5.69a), the ℓ_2–scaling constraint was employed to make the variance of every state equals to the variance of the input $y^*(k)$. Note that the inputs of realization (5.69b) are $u^*(k)$ and $y^*(k)$, and we found that for each coefficient wordlength B_c the variance of $u^*(k)$ is larger than the variance of $y^*(k)$. Therefore, for realization (5.69b), the ℓ_2–scaling constraint was employed to make the variance of every state equals to the variance of the input $u^*(k)$. The results are presented in table 5.5. The notation u in table 5.5 means the quadratic cost J^* is not well defined since the resulting closed–loop system is no longer stable due to coefficient quantization for a certain coefficient wordlength B_c (ie. $B_c \leq 12$).

□□□

In table 5.5, it can be seen that the degradation $\Delta\hat{J}$ for each coefficient wordlength B_c and for every compensator realization is not in a good agreement with the actual degradation ΔJ. In [Sasahara et. al (1984)], it is concluded that the structure which minimizes the effects of round–off noise (due to state quantization) also minimizes the effects of coefficient quantization. Intuitively, the same conclusion can be drawn by comparing the term tr(WZ) defined in (5.82a) and the sub–cost tr(WY) defined in (5.34a) where

$$Y = (\Phi_a - \varphi_a) I_a (\Phi_a - \varphi_a)'$$

where I_a is defined by (5.23d). To date, we have not been able to justify the conclusion analytically. However, the following example shows that the optimal FSWL structure derived using algorithm 5.3 also gives lower coefficient sensitivity.

Example 5.6 Consider again the minimal 6^{th}–order model used in example 5.1–5 where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider the quadratic cost J^* defined by (5.9a) where the weighting factors Q, M and R are respectively given by (A.23), (A.34) and (A.25). Consider also the optimum $\{K_q, G_q, T\}$ derived in example 5.4 for fixed state wordlength B=8.

Finite coefficient wordlength implementation of the FSWL compensator (5.19) gives

$$\hat{z}^*(k+1) = \tilde{\Phi}_2^*(\hat{z}^*(k)-\epsilon(k)) + \tilde{\varphi}\epsilon(k) + \tilde{K}^*\tilde{y}_e^*(k) \qquad (5.87a)$$

$$u^*(k) = -\tilde{G}^*(\hat{z}^*(k)-\epsilon(k)) + \delta(k) \qquad (5.87b)$$

where the quantized matrices $\tilde{\Phi}_2^*$, $\tilde{K}^*$ and $\tilde{G}^*$ are respectively the transformed version of the matrices Φ_2^*, K^* and G^* defined in (5.71c), (5.71d) and (5.71e), the residues $\epsilon(k)$ and $\delta(k)$ are defined in (5.36b-c), and the input $\tilde{y}_e^*(k)$ is defined in (5.36d). The residue matrix $\tilde{\varphi}$ is an $(n_x x n_x)$ identity matrix.

For fixed state wordlength B=8, the cost $\tilde{J}^*$ defined in (5.37a) was computed for the realization (5.87a-b) using the optimum $\{K_q,G_q,T\}$, and for some values of coefficient wordlength B_c. As a comparison, we also computed the cost $\tilde{J}^*$ for the triplet $\{K_\infty,G_\infty,T=I$ (ie. default structure)$\}$ where the gains K_∞ and G_∞ are given in (A.28) and (A.29). The results are presented in table 5.6.

B_c	$\tilde{J}^*(10^{-5})$	
	$\{K_\infty,G_\infty,T=I\}$	$\{K_q,G_q,T\}_{opt}$
18	9.7667	9.3124
16	9.7913	9.3276
14	9.8326	9.3403
12	9.8981	9.3916
10	u	9.4867
8	u	u

Table 5.6 The FWL performance resulting form
a. Ideal (q=0) design (denoted by $\{K_\infty,G_\infty,T=I\}$)
b. FSWL-LQG design (denoted by $\{K_q,G_q,T\}_{opt}$)

5.7 CONCLUSIONS

We have addressed the finite wordlength issue in designing an LQG regulator. The finite state wordlength (or round–off) noise affects

a. The selection of the Kalman filter and the controller gains

b. The choice of the compensator structure.

The optimum FWL–LQG design requires simultaneous optimization of the Kalman filter gain, the controller gain and the compensator structure. We first attacked the problem of designing the Kalman filter and the controller gains by assuming that the compensator structure is fixed (ie. the default structure). We have justified this assumption. The *integer residue correction* (IRC) scheme has been shown to improve the compensator performance. We have illustrated that a significant improvement can be achieved by taking the round–off noise into consideration in designing the Kalman filter and the controller gains. Unfortunately, the separation principle no longer applies. To date, no algebraic Riccati techniques have been found to aid the finite state wordelngth (FSWL) design solution. However, numerical examples show that the speed of convergence and the probability of obtaining the (possibly) global optimum can be improved by using algorithm 5.1.

The ideal Kalman filter and controller gains minimizes the ideal (ie. infinite precision) quadratic cost J in (5.3) but not the the finite wordlength (FWL) performance index J^* defined in (5.22a). Algorithm 5.1 produces the higher $\mathrm{tr}(\Upsilon P_1)$ but the resulting sub–cost $\mathrm{qtr}(\tilde{\Upsilon}\tilde{P}_2)$ is smaller for each given state wordlength B (see table 5.2). We have shown by means of examples that the FSWL–LQG design which can be solved iteratively using algorithm 5.3 gives considerable improvement over the ideal design particularly for a short state wordlength (ie. a small B in (5.17)). For example, for B=4 in table 5.4, the optimal FSWL–LQG design gives about 13% improvement on the performance index over the infinite precision design.

We have illustrated in section 5.6 that the finite coefficient wordlength plays an important role in the compensator implementation. Inappropriate selection of coefficient wordlength may lead to instability. For example, in table 5.5, it can be

seen that the three possible implementation (in (5.69a–c)) are all unstable when the coefficient wordlength $B_c \leq 12$.

By means of an example (ie. example 5.6), we have shown that the structure which minimizes the round–off noise (due to state quantization) also reduces the effects of coefficient quantization. In table 5.6, it can be seen that the ideal (q=0) design is unstable for coefficient wordlength $B_c \leq 10$ while the optimum FSWL design is unstable for $B_c \leq 8$. The two bits difference reflects the low sensitivity property of the optimum structure.

CHAPTER 6

A LOW COMPLEXITY - LOW SENSITIVITY COMPENSATOR IMPLEMENTATION

6.1 INTRODUCTION

In the implementation of a finite wordlength compensator, there are two main objectives which are commonly considered, namely the hardware complexity and the compensator performance. The hardware complexity of a FWL compensator is usually measured by the number of multiplications required per output sample while the FWL compensator performance is represented by the linear quadratic Gaussian (LQG) cost J. The default structure (which has been examined in sections 5.3) is usually derived directly from the physical model of a particular plant so that each state has a physical meaning. If the structure of the compensator is restricted to this default structure, then the hardware complexity (or structural issue) of the compensator is not necessary to be considered. A similarity transformation will alter the compensator state. Therefore, each of the transformed state does not represent the physical variable as originally required. However, if there is no restriction on the compensator structure then the FWL compensator is required to have a least complex structure (ie. minimum number of multiplications per output sample) and a minimum performance index J. Unfortunately, these two objectives can not be achieved simultaneously. For example, the optimum FSWL-LQG regulator described in section 5.5 has a minimum quadratic cost J but this optimum realization suffers from the fact that it has the most complex structure (ie. maximum number of multiplications per output sample) since in general all coefficients are neither zero nor unity.

The simplest structure to implement is a direct (canonic) realization [Phillips

and Nagle (1984), Oppenheim and Schafer (1975)]. There are many type of direct form structures, the most common types are the controllable and the observable canonical forms [Kailath (1980), Kuo (1980)]. Despite its least complex structure, a direct form structure exhibits some undersirable properties. The first drawback is that the non-zero and non-unity coefficients can be spread over a very large dynamic range [Hwang (1975b)]. Secondly, direct form structures produce a high round-off noise (and also a high coefficient sensitivity). This issue has been examined in detail particularly in digital filtering [Chan et. al (1973), Fettweis (1972), Jackson (1976)]. The last undesirable characteristics of a canonic realization is that it may have a high root (or pole-zero) sensitivity [Oppenheim and Schafer (1975), Williamson (1986)]. These disadvantages are not severe for the second order direct form structures [Barnes (1984a-b), Jackson et. al (1979), Lo and Jenq (1982)]. A higher order structure can be built using a cascade and/or parallel construction of 2^{nd}-order direct form structures. Using a series construction of 2^{nd}-order direct form structures, it is possible to optimally distribute the pole-zero pairs which can be followed by an internal block optimization [Hwang (1974), Jing and Fam (1986)].

In control applications, a series construction of 2^{nd}-order direct form structures is suitable for a single-input single-output (SISO) compensator. However, if the computational delay [Kwakernaak and Sivan (1972), Åström and Wittenmark (1984)] is critical then the cascade structure which requires a longer computation time is unfavorable. A parallel construction of 2^{nd}-order direct form structures [Jackson (1970), Willsky (1979)] is a good candidate as a multi-input multi-output (MIMO) compensator.

The high sensitivity nature of high order (>2) direct form structures can be improved by introducing the coefficient residue correction [Williamson and Sridharan (1986)] while the high round-off noise characteristic can be reduced by introducing the integer residue correction [Williamson and Sridharan (1985b)] or the sub-optimal error spectrum shaping [Abu-El-Haija and Peterson (1979)]. Another approach that can be utilized for improving the undersirable properties of high order direct form

structures is to use the so called *delay replacement* method. There are various types of delay replacement but basically all of them are derived using the following transformation

$$Z^{-1} = \frac{\gamma z^{-1}}{1 + \delta z^{-1}} \qquad (6.1)$$

which means each shift operator Z^{-1} in the original direct form structure is replaced by the feedback structure illustrated in Fig.6.1. An example of such transformation is shown in Fig.6.2 for a 2nd–order structure.

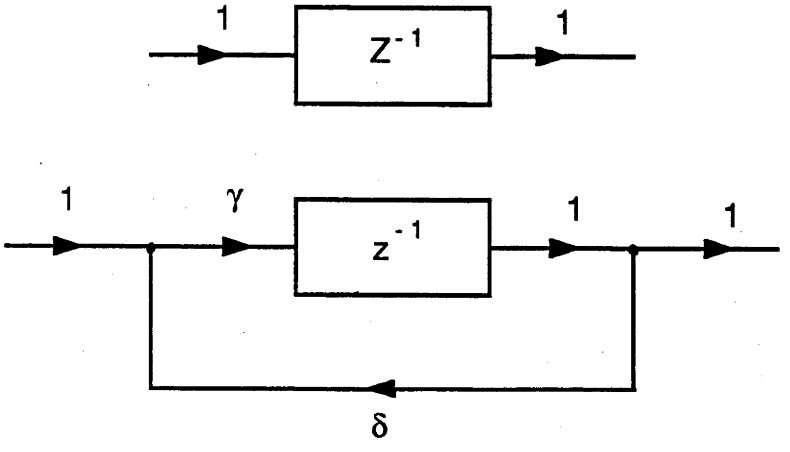
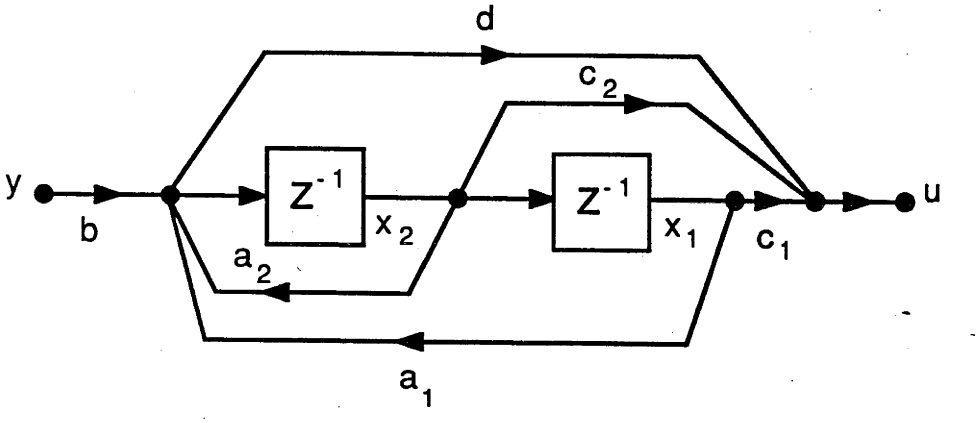


Fig. 6.1 Delay replacement transformation

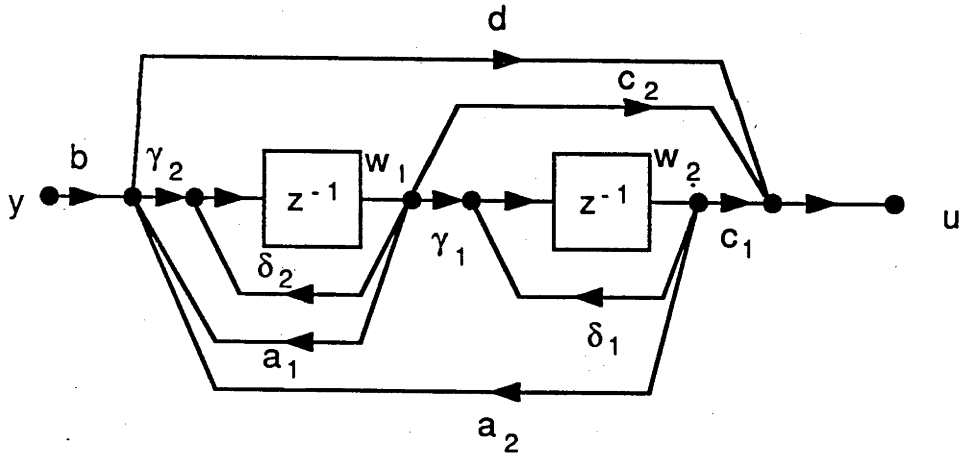
In digital filtering, various types of delay replaced direct form (DRDF) which correspond to different values of γ and δ in (6.1) have been proposed [Agarwal and Burrus (1975), Szczupak and Mitra (1978), Nishimura et. al (1981), Orlandi and Martinelli(1984)]. In control applications, the delay replacement idea has also been investigated. For example, in [Middleton and Goodwin (1986)] the delta operator which corresponds to $\gamma=\delta=1$ in (6.1) has been used to improve the deterministic characteristics of digitally controlled plants. Recently, a DRDF structure which has low complexity and low sensitivity properties is proposed in [Williamson (1987)]. The method which is termed the *scaled DRDF* is based on

$$Z^{-1} = \frac{\gamma_k z^{-1}}{1 + z^{-1}} \qquad (6.2)$$

that is to replace the k^{th} delay operator Z^{-1} in the original direct form structure by (6.2) where the factors γ_k are selected to satisfy some ℓ_2–scaling constraint.



(a)



(b)

Fig.6.2 2nd–order realizations of
a. a direct form structure, and
b. a delay replaced direct form structure

In this chapter, we consider the compensators which are implemented using the scaled DRDF structures. In section 6.2, we examine the use of single–input single–output scaled DRDF structures. A scaled DRDF structure can be derived from the given default structure. First, the default structure is transformed into a desired

controllable or observable canonical (or direct) structure. Then, by applying the frequency transformation (6.2) the scaled DRDF structure can be obtained. The performance of the compensator implemented using a scaled DRDF structure which requires only $3n_x$ multiplications (ie. the multiplications of the non-zero non-unity coefficients with each state and with the input) per output sample where n_x is the order of the compensator is compared to the performance of compensators implemented using a default structure and the optimal structure which requires $n_x^2+2n_x$ multiplications described in theorem 5.1 of section 5.4. The performance is measured in terms of the round-off noise (ie. due to state quantization) and the coefficient sensitivity (ie. due to coefficient quantization). The delay replacement transformation which is derived for SISO models can be generalized such that it can be used to transform a given MIMO direct form [Wolovich and Falb (1969), Luenberger (1967), Kailath (1980)] structure into a certain scaled MIMO-DRDF realization. This issue is investigated in section 6.3. The state quantization noise and the coefficient sensitivity performance of a compensator implemented using a certain MIMO-DRDF structure is also examined in section 6.3.

6.2 SINGLE INPUT SINGLE OUTPUT DELAY REPLACED DIRECT FORM COMPENSATOR STRUCTURES

In this section, we investigate the implementation of compensators using the scaled DRDF structure which will be derived from the given default structure. First, the effects of state quantization noise alone are examined. Both, the round-off and coefficient quantization noises are investigated afterwards.

Consider the minimal discrete-time model described by

$$x(k+1) = \Phi x(k) + \Gamma u(k) + \omega(k) \qquad (6.3a)$$

$$y(k) = Lx(k) + \eta(k) \qquad (6.3b)$$

where $x(k)$ is a simplified notation of $x(kT_c)$ for a sampling period T_c. The dimensions of the state, the input and the output of the model (6.3a-b) are

$$x(k) \epsilon R^{n_x} \quad ; \quad u(k) \epsilon R^{n_u} \quad ; \quad y(k) \epsilon R^{n_y}$$

The discrete process $\{\omega(k)\}$ and $\{\eta(k)\}$ are zero-mean wide sense stationary (WSS) processes having covariance

$$\xi\{\begin{bmatrix} \omega(k) \\ \eta(k) \end{bmatrix}[\omega'(k) \ \eta'(k)]\} = \begin{bmatrix} \Omega & 0 \\ 0 & \Lambda \end{bmatrix} \tag{6.4}$$

The infinite horizon performance index which we seek to minimize is described by

$$J = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} [x'(k)Qx(k) + 2x'(k)Mu(k) + u'(k)Ru(k)]\} \tag{6.5}$$

where the weighting matrices $Q \geq 0$, M and $R > 0$ are respectively $(n_x x n_x)$, $(n_x x n_u)$ and $(n_u x n_u)$. In an ideal situation (ie. infinite precision wordlength) the optimal compensator which minimizes the performance index (6.5) subject to the plant (6.3a-b) is governed by

$$\hat{x}(k+1) = \Phi\hat{x}(k) + \Gamma u(k) + K(y(k) - L\hat{x}(k)) \tag{6.6a}$$

$$u(k) = -G\hat{x}(k) \tag{6.6b}$$

where $\hat{x}(k)$ denotes the one-step-ahead state prediction $\hat{x}(k|y(k-1))$ and where K and G are respectively the Kalman filter and the controller gains. The optimal (stabilizing) gains K and G and the corresponding optimal quadratic cost J can be computed by means of lemma 5.2 of chapter 5.

The transformation described in following lemma can be used to transform a given (default) structure as in (6.3a-b) into some canonical realizations.

Lemma 6.1 [Chen (1984), Kuo (1980)] Consider the minimal discrete-time realization (6.3a-b). Let the charateristic polynomial of the system matrix Φ defined in (6.3a) be

$$\det(\lambda I_x - \Phi) = \lambda^{n_x} + \alpha_{n_x}\lambda^{n_x-1} + \ldots + \alpha_2\lambda + \alpha_1 \tag{6.7}$$

where det(.) denotes the determinant of a matrix. Define the controllability and observability matrices

$$C \triangleq [\Gamma \ \Phi\Gamma \ \ldots \ \Phi^{n_x-1}\Gamma] \tag{6.8a}$$

$$O \triangleq \begin{bmatrix} L \\ L\Phi \\ \\ \vdots \\ L\Phi^{n_x-1} \end{bmatrix} \tag{6.8b}$$

Then, there exist an equivalent transformation matrix which is non-singular

a. $T_c=CV$ for a single-input system (ie. $n_u=1$), such that the default structure $\{\Phi,\Gamma,L\}$ (6.3a-b) is transformed into the controllable canonical form

$$T_c^{-1}\Phi T_c \triangleq \Phi_c = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -\alpha_1 & -\alpha_2 & -\alpha_3 & \ldots & -\alpha_{n_x-1} & -\alpha_{n_x} \end{bmatrix} \tag{6.9a}$$

$$(T_c^{-1}\Gamma)' \triangleq \Gamma_c' = [0 \quad 0 \quad \ldots \quad 0 \quad 1] \tag{6.9b}$$

$$LT_c \triangleq L_c = [\beta_1 \quad \beta_2 \quad \ldots \quad \beta_{n_x-1} \quad \beta_{n_x}] \tag{6.9c}$$

where the matrix V is given by

$$V = \begin{bmatrix} \alpha_2 & \alpha_3 & \ldots & \alpha_{n_x} & 1 \\ \alpha_3 & \alpha_4 & \ldots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \alpha_{n_x} & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{bmatrix} \tag{6.10}$$

b. $T_0=VO$ for a single-output system (ie. $n_y=1$) such that the default structure $\{\Phi,\Gamma,L\}$ (6.3a-b) is transformed into the observable canonical form

$$T_0^{-1}\Phi T_0 \triangleq \Phi_0 = \Phi_c' \tag{6.11a}$$

$$T_0^{-1}\Gamma \triangleq \Gamma_0 = L_c' \tag{6.11b}$$

$$LT_0 \triangleq L_0 = \Gamma_c' \tag{6.11c}$$

where $\{\Phi_c,\Gamma_c,L_c\}$ is given by (6.9a–c) and V is defined by (6.10).

□□□

From an implementation point of view, the compensator (6.6a) can also be realized (and be implemented) as follows.

$$\hat{x}(k+1) = (\Phi-\Gamma G)\hat{x}(k) + K(y(k)-L\hat{x}(k)) \tag{6.12a}$$

$$\hat{x}(k+1) = (\Phi-KL)\hat{x}(k) + \Gamma u(k) + Ky(k) \tag{6.12b}$$

The advantage of the realization (6.12a) is that if the pair $\{\Phi,G\}$ is in a controllable canonical form (6.9a–b) then, the compensator matrix $(\Phi-\Gamma G)$ defined in (6.9a) is also in a controllable canonical form. The observable canonical form for the pair $\{\Phi,L\}$ as in (6.11a) and (6.11c) can be preserved if the compensator is implemented using realization (6.12b). In the subsequent analysis, we use the controllable canonical form. Therefore, the compensator that we consider is described by (6.12a). Another advantage of representation (6.12a) is that it has only one input (ie. $y(k)-L\hat{x}(k)$). Furthermore, it is well known that this input (which is known as the innovations sequence [Anderson and Moore (1979)]) is 'white'. These two facts are important in selecting the scaling procedure which is aimed at reducing the occurrence of the (fixed-point arithmetic) overflow. We will discuss the scaling method later in the chapter. In this section, we restrict the investigation to the single-input single-output (SISO) representation; that is the input u(k) and the output y(k) in (6.3a–b) are both scalar signals (and the dimensions $n_u=n_y=1$). The multi-input and multi-output (MIMO) representations will be examined in the next section.

The direct form (6.9a–c) can be transformed into a DRDF structure by using the frequency transformation (6.1) for $\gamma=\delta=1$. The result is stated in the following lemma.

Lemma 6.2 [Williamson (1987)] Consider the minimal discrete-time model (6.3a-b) where the system matrices $\{\Phi,\Gamma,L\}=\{\Phi_c,\Gamma_c,L_c\}$ are in the controllable canonical realization described in (6.9a-c). Then, applying the frequency transformation (6.1) for $\gamma=\delta=1$ to $\{\Phi_c,\Gamma_c,L_c\}$ results in the following delay replaced direct form (DRDF) realization.

$$T_1^{-1}\Phi_c T_1 \triangleq \Phi_1 = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ -\tilde{\alpha}_1 & -\tilde{\alpha}_2 & -\tilde{\alpha}_3 & \cdots & -\tilde{\alpha}_{n_x-1} & -\tilde{\alpha}_{n_x}+1 \end{bmatrix} \tag{6.13a}$$

$$(T_1^{-1}\Gamma_c)' \triangleq \Gamma_1' = \Gamma_c' \tag{6.13b}$$

$$L_c T_1 \triangleq L_1 = [\tilde{\beta}_1 \ \tilde{\beta}_2 \ \cdots \ \tilde{\beta}_{n_x-1} \ \tilde{\beta}_{n_x}] \tag{6.13c}$$

where

$$T_1(i,j) = \begin{cases} {}_{j-1}C^{i-j} & \text{if } i>j>1 \\ 0 & \text{if } i<j \end{cases} \tag{6.14a}$$

for all $i\epsilon[1,n_x]$ and $j\epsilon[1,n_x]$ and where

$${}_nC_r = \frac{n!}{(n-r)!\ r!} \tag{6.14b}$$

where n! denotes n factorial. Furthermore, the new coefficients in (6.13a) and (6.13c) are related to the direct form coefficients α_i and β_i for all $i\epsilon[1,n_x]$ in (6.9a) and (6.9c) by

$$\tilde{\alpha}_{n_x} = \alpha_{n_x} + 1 + T_1(n_x,n_x-1) \tag{6.15a}$$

$$\tilde{\alpha}_1 = \alpha_1 + T_1(n_x,1) + \sum_{j=1}^{n_x-1} T_1(j+1,1)\alpha_{j+1} \tag{6.15b}$$

$$\tilde{\alpha}_k = \alpha_k + \sum_{j=1}^{n_x-k} T_1(k+j,k)\alpha_{p+j} + T_1(n_x,k) + T_1(n_x,k-1) \qquad (6.15c)$$

for all k in the interval $1<k\leq n_x-1$, where $T_1(i,j)$ is given by (6.14a) and

$$\tilde{\beta}_{n_x} = \beta_{n_x} \qquad (6.15d)$$

$$\tilde{\beta}_k = \beta_k + \sum_{j=k+1}^{n_x} T_1(j+k-1,k)\beta_j \qquad (6.15e)$$

□□□

It has been shown in [Williamson (1987)] that the non-zero and non-unity coefficient (ie. $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ for all $i\epsilon[1,n_x]$) of the DRDF realization (6.13a-c) can be spread over a large dynamic range. One remedy to this particular problem is to allow the factors $\gamma_k\neq 1$ and let $\delta_k=1$ for all $k\epsilon[1,n_x]$ in the frequency transformation (6.1). The resulting structure is called the *scaled* DRDF structure.

Theorem 6.1 [Williamson (1987)] Consider the minimal discrete-time realization (6.3a-b) where the matrices $\{\Phi,\Gamma,L\}=\{\Phi_c,\Gamma_c,L_c\}$ are in the controllable canonical realization defined in (6.9a-c). Consider also the corresponding DRDF structure described by (6.13a-c), (6.14a-b) and (6.15a-e). Apply the frequency transformation (6.1) for $\gamma_k\neq 1$ and $\delta_k=1$ for all $k\epsilon[1,n_x]$ to the controllable canonical direct form $\{\Phi_c,\Gamma_c,L_c\}$. Then, the resulting DRDF structure is given by

$$T_2^{-1}\Phi_c T_2 \triangleq \Phi_2 = \begin{bmatrix} 1 & \gamma_1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \gamma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \gamma_{n_x-1} \\ -r_1 & -r_2 & -r_3 & \cdots & -r_{n_x-1} & -r_{n_x}+1 \end{bmatrix} \qquad (6.16a)$$

$$(T_2^{-1}\Gamma_c)' \triangleq \Gamma_2' = \Gamma_c' \qquad (6.16b)$$

$$L_c T_2 \triangleq L_2 = [s_1\tilde{\beta}_1 \quad s_2\tilde{\beta}_2 \quad \cdots \quad s_{n_x-1}\tilde{\beta}_{n_x-1} \quad s_{n_x}\tilde{\beta}_{n_x}] \qquad (6.16c)$$

where for all $m\epsilon[1,n_x]$

$$r_m = \tilde{\alpha}_m s_m \gamma_{n_x} \tag{6.17a}$$

$$s_{n_x} = \gamma_{n_x}^{-1} \tag{6.17b}$$

$$s_k = s\left(\prod_{j=k}^{n_x-1} \gamma_j\right)^{-1} \tag{6.17c}$$

for all k in the interval $1 \leq k \leq n_x - 1$ and where the coefficient $\tilde{\alpha}_m$ and $\tilde{\beta}_m$ are defined in (6.15c–e). Moreover, the equivalent transformation matrix T_2 is related to the transformation matrix T_1 defined in (6.14a) by

$$T_2 = T_1 S \tag{6.18a}$$

where S is a diagonal scaling matrix

$$S = \begin{bmatrix} s_1 & 0 & \cdots & 0 & 0 \\ 0 & s_2 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n_x-1} & 0 \\ 0 & 0 & \cdots & 0 & s_{n_x} \end{bmatrix} \tag{6.18b}$$

where the diagonal elements s_i for all $i \in [1, n_x]$ are given by (6.17b–c).

□□□

Substitution of T_2 in (6.16a) using (6.18a) gives

$$S^{-1} T_1^{-1} \Phi_c T_1 S = \Phi_2$$

and from (6.13a), we then have

$$S^{-1} \Phi_1 S = \Phi_2 \tag{6.19a}$$

Similarly, from (6.16c), we obtain

$$L_1 S = L_2 \tag{6.19b}$$

Therefore, the realization $\{\Phi_2, \Gamma_2, L_2\}$ defined in (6.16a–c) is nothing but the scaled version of the DRDF representation $\{\Phi_1, \Gamma_1, L_1\}$ defined in (6.13a–c). In terms of structural complexity, it can be seen from (6.16a–c) and (6.9a–c) that the scaled DRDF structure only requires n_x more multiplications per output sample than what

is required for the canonical structure. Note that the controllable canonical direct form structure (6.9a–c) requires $2n_x$ multiplications per output sample.

Suppose the state of the compensator (6.12a) are implemented using a finite fractional wordlength B (therefore, each state is less than unity), while all coefficients are presented using an arbritarily large (but finite) wordlength B_c. The finite (fixed point) state wordlength (FSWL) version of the compensator (6.12a) can be written as

$$\hat{x}^*(k+1) = (\Phi-\Gamma G)Q[\hat{x}^*(k)] + \varphi e(k) + K(y^*(k) - LQ[\hat{x}^*(k)]) \qquad (6.20a)$$

where φ is the residue feedback [Williamson and Sridharan (1985b)] matrix which is restricted to be an identity matrix (which has been justified in section 5.3 of the previous chapter) and e(k) is the state quantization residue defined below. The corresponding control law is governed by

$$u^*(k) = -GQ[\hat{x}^*(k)] + d(k) \qquad (6.20b)$$

where

$$Q[\hat{x}^*(k)] = \hat{x}^*(k) - e(k) \qquad (6.20c)$$

where e(k) and d(k) are the state quantization residue, and can be modelled [Sripad and Snyder (1977), Barnes et. al (1985)] as zero-mean 'white' noise processes with covariance qI_x where I_x is an $(n_x x n_x)$ indentity matrix and qI_u where I_u is an $(n_u x n_u)$ identity matrix, and q is given by

$$q = \frac{1}{12} 2^{-2B} \qquad (6.20d)$$

The discrete-time plant controlled by the FSWL compensator (6.20a) can be represented as follows.

$$x^*(k+1) = \Phi x^*(k) + \Gamma u^*(k) + \omega(k) \qquad (6.21a)$$

$$y^*(k) = Lx^*(k) + \eta(k) \qquad (6.21b)$$

Consequently, the quadratic cost to be optimized differs from the infinite precision performance index J defined in (6.5), and is described by

$$J^* = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} [x^*(k)'Qx^*(k) + 2x^*(k)'Mu^*(k) + u^*(k)'Ru^*(k)]\} \qquad (6.22)$$

In lemma 5.1 of the previous chapter, the quadratic cost J^* in (6.22) can be rewritten as

$$J^* = tr(\Upsilon P_1) + q[tr(\Upsilon P_2) + tr(R)] \qquad (6.23a)$$

where the covariance matrices P_1 and P_2 are defined to be

$$\xi\{\begin{bmatrix}\epsilon^*(k)\\ \hat{x}^*(k)\end{bmatrix}[\epsilon^*(k)' \ \hat{x}^*(k)']\} \triangleq P = P_1 + qP_2 \qquad (6.23b)$$

where q is defined in (6.20d) and $\epsilon^*(k)$ is the prediction error defined as

$$\epsilon^*(k) \triangleq x^*(k) - \hat{x}^*(k) \qquad (6.23c)$$

and where

$$\Upsilon = \begin{bmatrix} Q & Q-MG \\ (Q-MG)' & Q-2MG+G'RG \end{bmatrix} \qquad (6.23d)$$

where Q, M and R are defined in (6.22) and G is the controller gain defined in (6.20b).

Apply an arbitrary non-singular similarity transformation T on the FSWL compensator (6.20a-b), we obtain

$$\hat{z}^*(k+1) = \tilde{\Phi}_2 Q[\hat{z}^*(k)] + \tilde{\varphi}\epsilon(k) + \tilde{K}\tilde{y}_e^*(k) \qquad (6.24a)$$

$$u^*(k) = -\tilde{G}Q[\hat{z}^*(k)] + \delta(k) \qquad (6.24b)$$

$$Q[\hat{z}^*(k)] = \hat{z}^*(k) - \epsilon(k) \qquad (6.24c)$$

where

$$\tilde{y}_e^*(k) = y^*(k) - \tilde{L}Q[\hat{z}^*(k)] \qquad (6.24d)$$

$$\tilde{\Phi}_2 = T^{-1}\Phi_2 T \ ; \ \tilde{K} = T^{-1}K \ ; \ G = GT \qquad (6.24e)$$

The quantization residues $\delta(k)$ and $\epsilon(k)$ are not the same as the quantization residues d(k) and e(k) defined in (6.20b-c) although they will exhibit the same statistical models due to rounding.

The ℓ_2-scaling constraint that we discussed in section 5.2 of the previous

chapter is to make the variance of each compensator state to be equal to the variance of the compensator input (ie. the innovations sequence $\tilde{y}_e^*(k)$ defined in (6.24d)) which can be scaled to unity by multiplying the input $\tilde{y}_e^*(k)$ by a scaling factor s defined by

$$s = \frac{1}{(\sigma_y^2)^{\frac{1}{2}}} \tag{6.25a}$$

where

$$\sigma_y^2 \triangleq \xi\{\tilde{y}_e^*(k)^2\} \tag{6.25b}$$

Therefore, the ℓ_2–scaling constraint that we seek to satisfy is

$$\xi\{\hat{z}_j^*(k)^2\} = 1 \tag{6.25c}$$

for all $j\epsilon[1,n_z]$ where $n_z=n_x$ is dimension of the state $\hat{z}^*(k)$. We call the constraint (6.26) the *unity* ℓ_2–scaling constraint.

As a consequence of multiplying the compensator input by the scaling factor s defined in (6.25a), the output $u^*(k)$ has to be divided by the scaling factor s in (6.25a) in order to preserve the closed–loop transfer function. This is the so called *input/output* scaling procedure. Recall that the structure of the system matrix Φ is preserved. Therefore, the non–singular transformation T which is aimed at achieving the unity ℓ_2–scaling (6.25c) is restricted to be diagonal; the diagonal scaling matrix S defined in (6.18b) with properly selected diagonal elements s_i for all $i\epsilon[1,n_x]$ is the required matrix. For high input–signal to quantization–noise ratio (ie. $q \ll 1$ in (6.20d)), the covariance P in (6.23b) can be approximated by the covariance P_1 (ie. the term qP_2 in (6.23b) can be neglected). In this situation, the diagonal elements s_i which will satisfy the unity ℓ_2–scaling constraint is given by

$$s_i = (P_{jj})^{-\frac{1}{2}}$$

for all $i\epsilon[1,n_x]$ where P_{jj} is the $(j,j)^{th}$–element of the matrix P for all $j\epsilon[n_x+i,n_x+i]$.

The tranformed quadratic cost J^* can be represented as follows.

$$\tilde{J}^* = tr(\Upsilon P_1) + q[tr(\tilde{\Upsilon}\tilde{P}_2) + tr(R)] \tag{6.26}$$

where $\tilde{\Upsilon}$ and $\tilde{P}_2$ are the transformed version of Υ and P_2 defined in (6.23d) and

(6.23b). The detail of this transformation can be found in section 5.3 of the previous chapter. From (6.24), it can be seen only the term $tr(\tilde{\Upsilon}\tilde{P}_2)$ of the quadratic cost $\tilde{J}^*$ is affected by an arbitrary non-singular transformation T. The following lemma establishes the sub-cost $tr(\tilde{\Upsilon}\tilde{P}_2)$ for the scaled DRDF realization.

Lemma 6.3 Consider the minimal discrete-time model (6.21a) and the corresponding finite state wordlength (FSWL) compensator (6.20a). Consider as well the FSWL quadratic cost (6.23a). Suppose the model (6.21a), the FSWL compensator (6.20a) and the cost (6.23a) are transformed into a scaled delay replaced direct form structure by the (non-singular) similarity transformation T defined by

$$T = T_2 T_0 \tag{6.27a}$$

where T_0 and T_2 are respectively defined by (6.11a-c) and (6.18a). Then, the sub-cost $tr(\tilde{\Upsilon}\tilde{P}_2)$ given in (6.26) can be rewritten as follows.

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(I_a' \Phi_{ao}' \tilde{W} \Phi_{ao} I_a) \tag{6.27b}$$

where

$$I_a' = [I_X \quad -I_X] \tag{6.27c}$$

where I_X is an $(n_X x n_X)$ identity matrix and $\tilde{W}$ is governed by

$$\tilde{W}\Phi_{ao} + \Phi_{ao}'\tilde{W} + \Phi_{ao}'\tilde{W}\Phi_{ao} = -\tilde{\Upsilon} \tag{6.27d}$$

$$\tilde{\Upsilon} = \tilde{T}'\Upsilon\tilde{T} \tag{6.27e}$$

where Υ is defined in (6.23d) and

$$\Phi_{ao} = \begin{bmatrix} \Phi_{20}-KL & 0 \\ KL & \Phi_{20}-\Gamma G \end{bmatrix} \tag{6.27f}$$

$$\Phi_{20} = T^{-1}\Phi T - I_X \tag{6.27g}$$

$$\tilde{T} = \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix} \tag{6.27h}$$

Proof : From lemma 5.3 in chapter 5, we have

$$tr(\tilde{\Upsilon}\tilde{P}_2) = tr(I_a'(\tilde{\Phi}_a - \varphi_a)'\tilde{W}(\Phi_a' - \varphi_a) I_a) \tag{6.28a}$$

where I_a is defined in (6.27c) and

$$\tilde{\Phi}_a = \begin{bmatrix} \tilde{\Phi}-\tilde{K}\tilde{L} & 0 \\ \tilde{K}L & \tilde{\Phi}-\tilde{\Gamma}\tilde{G} \end{bmatrix} \tag{6.28b}$$

$$\tilde{\Phi} = T^{-1}\Phi T \quad ; \quad \tilde{K} = T^{-1}K \tag{6.28c}$$

$$\tilde{\Gamma} = T^{-1}\Gamma \quad ; \quad \tilde{L} = LT \quad ; \quad \tilde{G} = T^{-1}G \tag{6.28d}$$

where T is defined in (6.27a), and where

$$\varphi_a = \begin{bmatrix} I_x & 0 \\ 0 & I_x \end{bmatrix} \tag{6.28e}$$

$$\tilde{W} = \tilde{\Phi}_a'\tilde{W}\tilde{\Phi}_a + \tilde{\Upsilon} \tag{6.28f}$$

where $\tilde{\Upsilon}$ is defined in (6.27e). However, the system matrix Φ which is equivalent to Φ_2 in (6.16a) can be decomposed as

$$\tilde{\Phi} = \Phi_2 = \Phi_{20} + I_x \tag{6.28g}$$

Therefore, $\tilde{\Phi}_a$ in (6.28b) can also be decomposed as

$$\tilde{\Phi}_a = \Phi_{ao} + \varphi_a \tag{6.28h}$$

where Φ_{ao} and φ_a are defined in (6.27f) and (6.28e) respectively. Substitution of $\tilde{\Phi}_a$ in (6.28a) using (6.28h) gives (6.27b). Substitute $\tilde{\Phi}_a$ in (6.28f) using (6.28h), after a little manipulation we obtain (6.27d).

□□□

In the following example, we illustrate the round-off noise performance of a compensator implemented using the scaled DRDF structure.

Example 6.1 Consider the minimal 6^{th} order discrete-time model (6.21a-b) where the system matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) in Appendix A, and where the covariance matrices Ω and Λ are given by (A.20) and (A.21) respectively. Consider also the quadratic cost J^* defined by (6.22) where the weighting matrices Q, M and R are given by (A.23), (A.24) and (A.25) respectively. The infinite precision gains K_∞ and G_∞ which minimizes J^* for q=0 (ie. B=∞ in (6.20d)) subject to the plant (6.21a-b) are respectively given by (A.28) and (A.29).

Apply the similarity transformation T defined in (6.27a) on the default structure

representation $\{\Phi,\Gamma,L\}$, $\{\Omega,\Lambda\}$ and $\{Q,M,R\}$ where the scaling matrix S defined in (6.18b) was selected such that every state of the transformed compensator is unity ℓ_2-scaled (ie. satisfies the unity ℓ_2-scaling (6.25c)). The resulting sub-cost $\mathrm{tr}(\tilde{T}\tilde{P}_2)$ defined in (6.27b) for each state wordlength B is presented in table 6.1. For the purpose of comparison, for each state wordlength B, the sub-cost $\mathrm{tr}(\tilde{T}\tilde{P}_2)$ was also calculated for the default and the controllable direct form structures with scaling and for the gains K_∞ and G_∞.

wordlength B	$\mathrm{qtr}(\tilde{T}\tilde{P}_2)$ (10^{-6})		
	default structure	scaled DRDF	scaled DF
10	0.2431	0.2465	0.2489
8	1.5183	1.5420	1.6131
6	3.2796	3.5616	4.7682
4	33.0387	37.1011	42.03229

Table 6.1 The round-off noise performance of compensators implemented using a
a. Default structure
b. Scaled DRDF structure
c. Direct form (DF) structure

Finite coefficient implementation of the FSWL compensator (6.24a-e) can be represented as

$$\hat{z}^*(k+1) = \tilde{\Phi}_2^* Q[\hat{z}^*(k)] + \tilde{\varphi}\epsilon(k) + \tilde{K}^* \tilde{y}_e^*(k) \qquad (6.29a)$$

$$u^*(k) = -\tilde{G}^* Q[\hat{z}^*(k)] + \delta(k) \qquad (6.29b)$$

where $Q[\hat{z}^*(k)]$ and $\tilde{y}_e^*(k)$ are defined in (6.24c-d), and the coefficient A^* is the quantized version of the ideal coefficient A to a B_c-bit fractional representation, and

$$\tilde{\Phi}_2^* = (T^{-1}\Phi_2 T)^* \quad ; \quad \tilde{K} = (T^{-1}K)^* \quad ; \quad \tilde{G} = (GT)^* \qquad (6.29c)$$

and where the quantization residues $\epsilon(k)$ and $\delta(k)$ are defined in (6.24b-c).

For fixed state wordlength B=8, the quadratic cost $\tilde{J}^*$ defined in (6.26) was computed for the compensator (6.29a-c) for some values of coefficient wordlength B_c, for the gains K_∞ and G_∞ given by (A.28) and (A.29) and for the default, the

controllable scaled DRDF, the optimum structure (derived by algorithm 5.2 of the previous chapter) and the controllable direct form structures. The results are presented in table 6.2. The notation u in table 6.2 means the cost $\tilde{J}^*$ is not well defined since the resulting closed-loop system is no longer stable due to coefficient quantization for a certain coefficient wordlength B_c.

B_c	$\tilde{J}^*(10^{-5})$			
	Defaults	DRDF	Optimum	DF
18	9.7667	9.7667	9.7665	9.7668
16	9.7672	9.7670	9.7667	9.7689
14	9.7695	9.7689	9.7678	9.7801
12	9.7798	9.7741	9.7699	u
10	u	9.7938	9.7782	u
8	u	u	u	u

Table 6.2 The FWL performance of compensators implemented using
a. Default structure (denoted by Default)
b. Controllable scaled DRDF structure (denoted by DRDF)
c. Optimum FSWL structure (denoted by optimum)
d. Controllable direct form structure (denoted by DF).

6.3 MULTI-INPUT MULTI-OUTPUT DELAY REPLACED DIRECT FORM COMPENSATOR STRUCTURES

The use of a single-input single-output (SISO) DRDF structure in the control aplication has been investigated in the previous section. It is shown by means of an example that a SISO-DRDF structure which has a low complexity property (which requires $3n_x$ number of multiplications per output sample) is less sensitive to the coefficient change which is due to the coefficient quantization. In this section, we examine the FWL performance of compensators implemented using *multi-input multi-output* (MIMO) DRDF structures.

So far, the scaling issue that has been addressed concerns only the SISO

representation. Recall the unity ℓ_2–scaling constraint (described in chapter 5) in which the probability of overflow in each state of the compensator is made equal to the probability of overflow of the compensator input which has been normalized to unity. If the realization (6.20a) is *single–input multi–output* (SIMO), then the SISO unity ℓ_2–scaling technique can obviously be used without causing any problem. However, for a *multi–input* compensator there is a problem. To see this consider the 2–input compensator depicted in Fig.6.3.

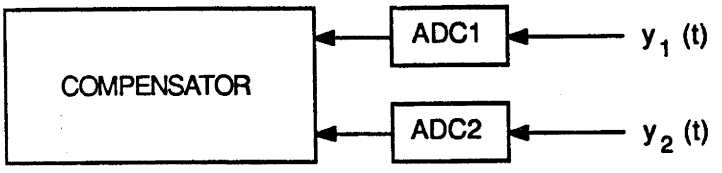


Fig.6.3 A 2–input digital compensator

In general, the variances of the compensator input (or the plant output) $y_1(t)$ and $y_2(t)$ are not the same. Therefore, the objective of equalizing the probability of overflow of each compensator state and of every compensator input can not be achieved unless the wordlengths of the ADC1 and of the ADC2 are allowed to be different [Moroney (1983)] assuming the scaling is done also with respect to the quantization noise. In this section, we avoid this difficulty by selecting only one of the compensator inputs (ie. the one which has maximum variance) to have the same probability of overflow (after scaling) as all compensator states and let the remaining of the compensator inputs to have a less probability of overflow. Using this approach, the SISO unity ℓ_2–scaling technique can be used but now the plant output $y_m(t)$ (which satisfies $\xi\{y_m^2(t)\} \geq \xi\{y_i^2(t)\}$ for all $i \in [1, n_y]$ where n_y is the dimension of the plant output y(t)) has to be selected.

Consider the minimal discrete–time MIMO (default structure) plant (6.3a–b) and the corresponding infinite precision compensator (6.12a). The default representation $\{\Phi,\Gamma,L\}$ can be transformed into some canonical observability or controllability forms. It has been shown that unlike the single variable (ie. SISO) case, the MIMO direct

form structures are generally not unique [Kailath (1980), Luenberger (1967), Wolovich and Falb (1969)]. However, the structure of the canonical form can be specified to some extent to meet some design specifications. A method of transforming a given controllable default structure into some controllability forms is established in the following lemma.

Lemma 6.4 [Luenberger (1967), Wolovich and Falb (1969)] Consider the minimal (default structure) realization $\{\Phi,\Gamma,L\}$ (6.6a-b) where the dimensions of the matrices Φ, Γ and L are respectively $(n_x x n_x)$, $(n_x x n_u)$ and $(n_y x n_x)$ and the controllability matrix C which is given by (6.8a). Suppose the matrix L which is defined by

$$L \triangleq [b_1, \Phi b_1, \ldots, \Phi^{\sigma_1 - 1} b_1, b_2, \Phi b_2, \ldots, \Phi^{\sigma_2 - 1} b_2, \ldots, b_{n_u}, \Phi b_{n_u}, \ldots, \Phi^{\sigma_{n_u} - 1} b_{n_u}] \tag{6.30}$$

where the $\{\sigma_i; 1 \leq i \leq n_u\}$ are the non-negative controllability indices and b_i for all $i \epsilon [1, n_u]$ denotes the i^{th} column of the input matrix Γ consists of n_x linearly independent columns of the controllability matrix C. Let

$$d_0 = 0 \quad ; \quad d_k = \sum_{i=1}^{k} \sigma_i \tag{6.31}$$

for all $k=1,2,\ldots n_u$. Suppose e_k be the d_k^{th} row of the inverse matrix L^{-1}.

Then, the following matrix

$$T_c = \begin{bmatrix} e_1 \\ e_1\Phi \\ \vdots \\ e_1\Phi^{\sigma_1-1} \\ e_2 \\ e_2\Phi \\ \vdots \\ e_2\Phi^{\sigma_2-1} \\ \vdots \\ e_{n_u} \\ e_{n_u}\Phi \\ \vdots \\ e_{n_u}\Phi^{\sigma_{n_u}-1} \end{bmatrix} \tag{6.32}$$

generates a Lyapunov transformation for which the transformed realization $\{T_c^{-1}\Phi T_c, T_c^{-1}\Gamma, LT_c\}$ is in a companion form. More precisely,

$$T_c^{-1}\Phi T_c \triangleq \Phi_c = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1n_u} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2n_u} \\ \vdots & \vdots & & \vdots \\ \Phi_{n_u1} & \Phi_{n_u2} & \cdots & \Phi_{n_un_u} \end{bmatrix} \tag{6.33}$$

with Φ_{ii} is a $(\sigma_i x \sigma_i)$ companion matrix defined by

$$\Phi_{ii} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \alpha_{d_i,d_{i-1}+1} & \alpha_{d_i,d_{i-1}+2} & \cdots & \alpha_{d_i,d_{i-1}} & \alpha_{d_i,d_i} \end{bmatrix} \tag{6.34}$$

and Φ_{ij} $(i \neq j)$ is a $(\sigma_i x \sigma_j)$ matrix given by

$$\Phi_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \alpha_{d_i, d_{j-1}+1} & \cdots & \alpha_{d_i, d_j} \end{bmatrix} \tag{6.35}$$

and

$$T_c^{-1}\Gamma \triangleq \Gamma_c = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & \Gamma_{d_1,2} & \Gamma_{d_1,3} & \cdots & \Gamma_{d_1,n_u} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \Gamma_{d_2,3} & \cdots & \Gamma_{d_2,n_u} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \} \sigma_1 \\ \\ \\ \} \sigma_2 \\ \\ \\ \} \sigma_{n_u} \end{matrix} \tag{6.36a}$$

where $\Gamma_{i,j}$ denotes the $(i,j)^{th}$ element of a matrix Γ, and

$$LT_c = L_c \tag{6.36b}$$

□□□

Note that the non-uniqueness of the MIMO direct form realization is due to the freedom in constructing the matrix L in (6.18) which is based on the factors σ_i. The only restriction in constructing the matrix L is that no vector of the form $\Phi^i b_j$ is selected unless the vector $\Phi^n b_j$ for $n<i$ are also selected. There are many ways of constructing the matrix L [Kailath (1980)]. For our purposes, we use a so called Young diagram [Kalman (1972)].

From (6.33-36), the number of multiplications required per output sample for a (controllable) n_u-input n_y-output n_x^{th}-order MIMO direct form structure is

$$n_u n_x + n_y n_x + \sum_{i=1}^{n_u-1} (n_u - i) \tag{6.36c}$$

Note that in deriving (6.36c) we only consider the non-zero and the non-unity coefficients in (6.33-36). For a SISO direct form structure, the number of multiplications required can also be expressed as (6.36c). But since $n_u=n_y=1$, the expression (6.36c) becomes $2n_x$.

A direct form SISO model can be transformed into a DRDF structure by means of lemma 6.2. A MIMO direct form structure derived in lemma 6.4 involves some companion forms as in (6.34). A generalization of the result in lemma 6.2 to the MIMO case can be done by transforming each companion form Φ_{ii} in (6.34) into a DRDF structure. We present the result in the following lemma.

Lemma 6.5 Consider the controllable direct form realization $\{\Phi_c,\Gamma_c,L_c\}$ described in (6.9a-c). Consider as well the frequency transformation (6.1) which produces the elements of the transformation (6.14a). Then, applying the equivalent transformation

$$T_1 = \begin{bmatrix} T_{11} & 0 & \dots & 0 \\ 0 & T_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & T_{n_u n_u} \end{bmatrix} \tag{6.37}$$

where $T_{kk}(i,j)$ for $k\epsilon(1,n_u)$ and for $i\epsilon[1,\sigma_k]$ and $j\epsilon[1,\sigma_k]$ (where σ_i for all $i\epsilon[1,k]$ are defined in (6.30)) is defined by (6.14a) to the direct form $\{\Phi_c,\Gamma_c,L_c\}$ gives the following DRDF structure.

$$T_1^{-1}\Phi T_1 \triangleq \Phi_1 = \begin{bmatrix} \tilde{\Phi}_{11} & \tilde{\Phi}_{12} & \dots & \tilde{\Phi}_{1n_u} \\ \tilde{\Phi}_{21} & \tilde{\Phi}_{22} & \dots & \tilde{\Phi}_{2n_u} \\ \vdots & \vdots & & \vdots \\ \tilde{\Phi}_{n_u 1} & \tilde{\Phi}_{n_u 2} & \dots & \tilde{\Phi}_{n_u n_u} \end{bmatrix} \tag{6.38a}$$

where $\tilde{\Phi}_{ii}$ is a $(\sigma_i x \sigma_i)$ matrix given by

$$\tilde{\Phi}_{ii} \triangleq T_{ii}^{-1}\Phi_{ii}T_{ii}$$

$$= \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 1 \\ \tilde{\alpha}_{d_i, d_{i-1}+1} & \tilde{\alpha}_{d_i, d_{i-1}+2} & \cdots & \tilde{\alpha}_{d_i, d_{i-1}} & \tilde{\alpha}_{d_i, d_i} \end{bmatrix} \tag{6.38b}$$

and $\tilde{\Phi}_{ij}$ is a $(\sigma_i x \sigma_j)$ matrix given by

$$\tilde{\Phi}_{ij} = T_{ii}^{-1}\Phi_{ij}T_{jj} = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \tilde{\alpha}_{d_i, d_{j-1}+1} & \cdots & \tilde{\alpha}_{d_i, d_j} \end{bmatrix} \tag{6.38c}$$

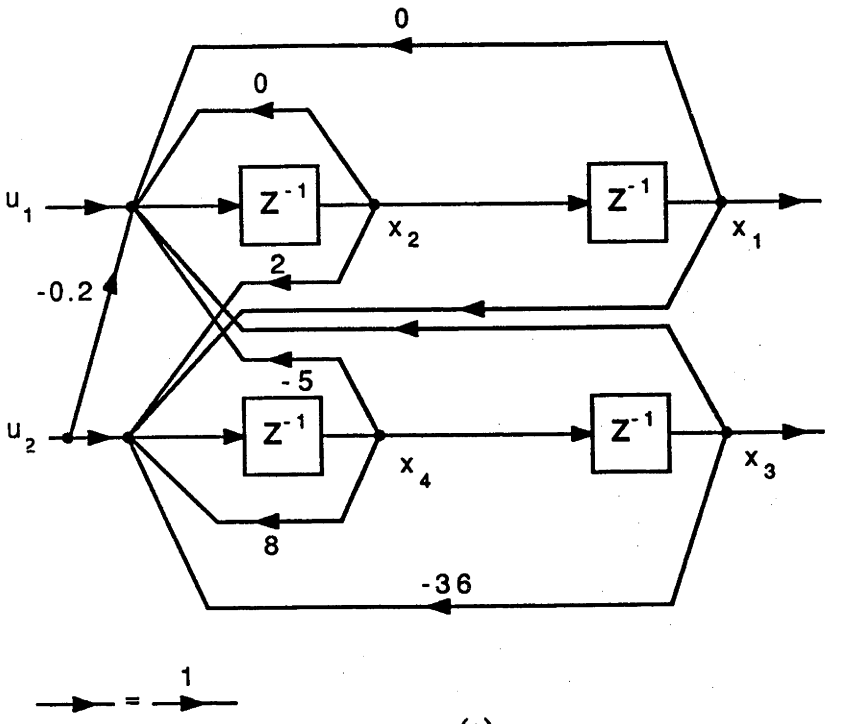
and where

$$T_1^{-1}\Gamma \triangleq \Gamma_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & \tilde{\Gamma}_{d_1,2} & \tilde{\Gamma}_{d_1,3} & \cdots & \tilde{\Gamma}_{d_1,n_u} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \tilde{\Gamma}_{d_2,3} & \cdots & \tilde{\Gamma}_{d_2,n_u} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \}\sigma_1 \\ \\ \\ \}\sigma_2 \\ \\ \\ \}\sigma_{n_u} \end{matrix} \tag{6.38d}$$

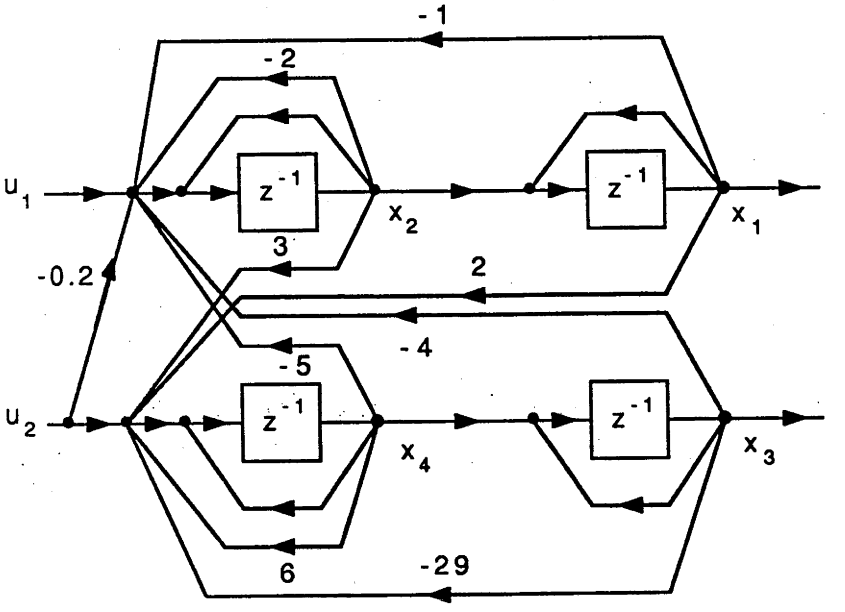
$$L_c T_1 \triangleq L_1 \tag{6.38e}$$

□□□

An example of a 2-input 2-output DRDF structure is depicted in Fig.6.4.



(a)



(b)

Fig.6.4 4th-order 2-input 2-output compensators realized using
a. direct form structure.
b. DRDF structure.

Similar to the single variable case, the non-zero and the non-unity coefficient of the MIMO-DRDF structure may be spread over a large dynamic range. A proper scaling is one solution to this problem. The following theorem which is an extension of theorem 6.1 establishes the construction of the scaled MIMO-DRDF structure.

Theorem 6.2 Consider the minimal MIMO default structure realization $\{\Phi,\Gamma,L\}$ defined in (6.3a-b) where the dimensions of the system matrices Φ, Γ and L are respectively given by $(n_x x n_x)$, $(n_x x n_u)$ and $(n_y x n_x)$. Consider as well the equivalent transformation matrices T_c defined in (6.32) and T_1 defined in (6.37) and the corresponding controllability form $\{\Phi_c,\Gamma_c,L_c\}$ defined by (6.33-36) and the DRDF structure defined by (6.38a-e). Suppose S is a scaling matrix of the form

$$S = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & S_{n_u} \end{bmatrix} \qquad (6.39a)$$

where s_i is a $(\sigma_i x \sigma_i)$ matrix given by

$$S_i = \begin{bmatrix} s_{i1} & 0 & \dots & 0 \\ 0 & s_{i2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_{id_i} \end{bmatrix} \qquad (6.39b)$$

where d_i is defined by (6.31) and the corresponding scalar σ_i is defined by (6.30) for all $i\epsilon[1,n_u]$. Then, the similarity transformation $T_2=ST_1T_c$ transforms the default structure $\{\Phi,\Gamma,L\}$ into a scaled DRDF structure of the form

$$T_2^{-1}\Phi T_2 \triangleq \Phi_2 = \begin{bmatrix} \bar{\Phi}_{11} & \bar{\Phi}_{12} & \cdots & \bar{\Phi}_{1n_u} \\ \bar{\Phi}_{21} & \bar{\Phi}_{22} & \cdots & \bar{\Phi}_{2n_u} \\ \vdots & \vdots & & \vdots \\ \bar{\Phi}_{n_u1} & \bar{\Phi}_{n_u2} & \cdots & \bar{\Phi}_{n_un_u} \end{bmatrix} \tag{6.40a}$$

where $\bar{\Phi}_{ii}$ is a $(\sigma_i x \sigma_i)$ companion matrix given by

$$\bar{\Phi}_{ii} = \begin{bmatrix} 1 & \gamma_{i1} & 0 & \cdots & 0 & 0 \\ 0 & 1 & \gamma_{i2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \gamma_{id_{i-1}} \\ \bar{\alpha}_{d_i,d_{i-1}+1} & \bar{\alpha}_{d_i,d_{i-1}+2} & \bar{\alpha}_{d_i,d_{i-1}+3} & \cdots & \bar{\alpha}_{d_i,d_{i-1}} & \bar{\alpha}_{d_i,d_i} \end{bmatrix} \tag{6.40b}$$

and Φ_{ij} is a $(\sigma_i x \sigma_i)$ matrix given by

$$\bar{\Phi}_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \bar{\alpha}_{d_i,d_{j-1}+1} & \cdots & \bar{\alpha}_{d_i,d_j} \end{bmatrix} \tag{6.40c}$$

$$T_2^{-1}\Gamma \triangleq \Gamma_2 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \gamma_{1d_1} & \gamma_{1d_1}\tilde{\Gamma}_{d_1,2} & \gamma_{1d_1}\tilde{\Gamma}_{d_1,3} & \cdots & \gamma_{1d_1}\tilde{\Gamma}_{d_1,n_u} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \gamma_{2d_2} & \gamma_{2d_2}\tilde{\Gamma}_{d_2,3} & \cdots & \gamma_{2d_2}\tilde{\Gamma}_{d_2,n_u} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \gamma_{n_u d_{n_u}} \end{bmatrix} \begin{matrix} \\ \\ \}\sigma_1 \\ \\ \}\sigma_2 \\ \\ \\ \}\sigma_{n_u} \end{matrix} \qquad (6.40d)$$

where $\tilde{\Gamma}_{ij}$ are the elements of the matrix Γ_1 defined in (6.38d) and

$$LT_2 \triangleq L_2 \qquad (6.40e)$$

□□□

From the controllable scaled DRDF matrices given in (6.40a–e), the number of multiplications (denoted by the number of non–zero and non–unity coefficients) required per output sample is

$$n_u n_x + n_x + n_y n_x + \sum_{i=1}^{n_u-1} (n_u - i) \qquad (6.40f)$$

Therefore, from (6.36c) and (6.40f) it can be deducted that the controllable scaled MIMO–DRDF structure requires only n_x more multiplications than would be required by the SISO direct form.

In the previous section, using theorem 5.2 the combine finite state and coefficient wordlength performance of a compensator implemented using the controllable scaled DRDF structure has been investigated by means of an illustrative example. In the following example, we examine the performance of a MIMO compensator implemented using the scaled DRDF structure and compare to the

performance of compensators implemented using different structures.

Example 6.3 Consider the 10^{th} order 2-input 2-output model of a 5-mode flexible (Euler-Bernolli) beam which can be described by realization (6.21a-b) for n_x=10 and n_u=n_y=2 and where the matrices Φ, Γ and L are given by (B.21), (B.22) and (B.23) respectively in appendix B and the covariances Ω and Λ are given by (B.25) and (B.26) respectively. Consider as well the quadratic cost J^* defined in (6.22) where the weighting matrices Q, M and R are respectively given by (B.28), (B.29) and (B.30).

The infinite precision Kalman filter and the controller gains K_∞ and G_∞ which minimizes the quadratic cost J^* in (6.22) for q=0 in (6.20d) subject to the plant (6.21a-b) are given by (B.33) and (B.34). The finite (state and coefficient) wordlength compensator in which the IRC term is incorporated is given by (6.29a-b). The scaled DRDF structure was derived by means of theorem 6.2 where the scaling matrix S defined in (6.39a) was selected such that every compensator state is unity ℓ_2-scaled with respect to the first input $(y_c^*(k))_1$.

wordlength B_c	$\tilde{J}^*(10^{-5})$		
	default	scaled DRDF	DF
20	4.9138	4.9883	5.1134
18	4.9169	5.7367	5.9232
16	5.6372	6.8240	u
14	u	u	u

Table 6.3 The FWL performance of MIMO compensators implemented using
a. Default structure
b. Controllable scaled DRDF structure
c. Controllable direct form (DF) structure

For fixed state wordlength B=10, the performance index J^* in (6.26) was computed for different value of coefficient wordlength B_c, for the compensator (6.29a-b) and

for the gains K_∞ and G_∞. As a comparison, we also computed the quadratic cost J^* for the default and the direct form structures where both structures are unity ℓ_2-scaled (with respect to the first input $(\hat{y}_e^*(k))_1$. The results are presented in table 6.3. The notation u in table 6.3 means the quadratic cost $\tilde{J}^*$ is not well defined since the resulting closed-loop system is no longer stable due to coefficient quantization for a certain coefficient wordlength B_c.

6.4 CONCLUSIONS

In this chapter, we have examined the compensator implementation using delay replaced direct form (DRDF) structures. The idea of using delay replacement in direct form is not new, it is very well known in digital filtering. The low complexity property of DRDF structure can be deduced directly from the corresponding representation while the low sensitivity to the coefficient change due to finite coefficient wordlength property has been illustrated by means of examples. We have considered the implementation of both the single-input single-output (SISO) and the multi-input multi-output (MIMO) DRDF structures. In both cases, we have developed the procedures which can be used to obtain a certain DRDF structure directly from the given default structure.

The low complexity property of the scaled DRDF structures can be shown by the number of multiplications required per output sample. For the scaled SISO-DRDF structure, the required number of multiplications is $3n_x$ where n_x is the dimension of the compensator state. This implies that the scaled SISO-DRDF structure requires n_x more multiplications than would be required by the direct form structure. For the (controllable) scaled MIMO-DRDF structure, the required number of multiplications required per output sample is

$$n_x n_u + n_x + n_x n_y + \sum_{i=1}^{n_u-1} (n_u - i)$$

where n_u and n_y are the dimension of the compensator input and output

respectively. As for the scaled SISO–DRDF structure, the scaled MIMO–DRDF structure also requires n_x more multiplications than would be required by the MIMO direct form structure. For the 2–input 2–output 10^{th}–order model used in example 6.3, the required number of multiplications per output sample is 41 for the direct form structure and is 51 for the scaled (controllable) DRDF structure. Note that the maximum number of multiplications required per output sample for this particular model is 140.

The FWL performance of a SISO–DRDF structure is compatible with the default structure as depicted in table 6.2 even though its FSWL performance is worst (see table 6.1). In fact, this example reflects the low sensitivity property of the DRDF structure.

In the flexible beam example, the FWL performance of the scaled MIMO–DRDF structure is not as good as the FWL performance of the default structure. This is due to the fact that the default structure is already in a special form (ie. block form) which is 'close' to the optimal form.

CHAPTER 7

SUMMARY, CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

The thesis concerns with the cyclostationary and the finite wordlength issues which are particularly relevant for the fast rate digital control systems. It can roughly be divided into two parts. The first part (chapters 2, 3 and 4) discusses the consequences of cyclostationary processes on the digital control design. Some methods of characterizing cyclostationary processes and improving intersample behavior of digitally controlled continuous-time systems is proposed and non-synchronous and multirate optimal state prediction are investigated. The second part (chapters 5 and 6) examines finite fixed-point wordlength nature of the implementation on the design of the Kalman filter and controller gains. The implementation of a low complexity - low sensitivity compensator is investigated.

In chapter 2, we considered the translational and the harmonic series representations of WSCS processes. In particular, we developed a state space translational representation of WSCS processes and used it to characterize two classes of first and second order processes. For a high damping factor ζ, the continuous-time characteristics of the covariance $\Omega(\delta)$ are similar to the curves of the covariance of first order systems for $w_n=2\beta$, where β and w_n are respectively the time constant and the natural frequency of the model (see Fig.2.3 and Fig.2.4). It is postulated that for an N^{th}-order overdamped system, the characteristics of the covariance function $\Omega(\delta)$ are described by $w_n=N\beta$. This approximation requires further investigation. A first order approximation of a high order harmonic series

representation examined in this chapter. Numerical results revealed (see Fig.2.4) that a first order $\hat{\Omega}(\delta)$ is a good approximation of $\Omega(\delta)$ for $w_n < \pi T_c^{-1}$ and $\zeta < 1$ where T_c is sampling period. This demands further justification.

In chapter 3, we examined the consequences of the cyclostationary process disturbance and measurement noise on the state prediction problem. When the period of the statistical properties of the cyclostationary disturbances and the control sampling period are equal, we showed that the optimal state prediction is not generally achieved by a synchronous control and output sampling. In both the single and multirate cases, investigation was carried out under the assumption that the process disturbance and the measurement noise are cyclostationary of period $T_\omega = T_\eta = T_c$. Generally speaking, the state prediction that we explored in this chapter is limited to a special class (ie. $T_\omega = T_\eta = T_c$) even though it is more general than the conventional state prediction problem.

In chapter 4, a quadratic cost function which takes the intersample characteristics into account was proposed and minimized. Specifically, we showed that the proposed minimax quadratic problem is a well defined problem; that is by showing that the controllability and the observability conditions are well preserved. Therefore, more results are needed in order to establish a similar result for the general case. In the case of output variance regulation, we showed that the MMVR can significantly improve the intersample behavior of the digitally controlled continuous-time systems. For the examples considered in chapter 4, the algorithm 4.1 converges after 2 updates (<5) of δ_i with initial $\delta_0 = 0.5$. The convergence of this algorithm requires an analytical justification. We were unable to illustrate the use of a more general MMVR even though theoretically it can be accomplished by including several internal weightings which corresponds to $N > 1$ in the quadratic cost (4.19a-b).

In chapter 5, we considered the finite wordlength LQG regulator design. We investigated the optimum fixed-point FWL-LQG regulator which involves the optimum Kalman filter and the optimum controller gains and the optimum

compensator structure. We showed that the finite state wordlength nature of the implementation should be taken into consideration when selecting the predictor and controller gains. The inclusion of the integer residue correction (IRC) scheme has been shown to improve the compensator performance. The main difficulty that we faced in approaching the finite state wordlength problem for a fixed structure (see FSWL design 1 of section 5.3) is that the state prediction and control problems are not separable. The FSWL Kalman filter design discussed in [Williamson (1985)] provides a nice expression for the optimal estimation gain. The FWL-MVR design which is a special case of the FWL-LQG design is more likely to be an issue that can be explored analytically. In particular for a single-input single-output system, we were able to obtain a sufficient condition for the compensator gain to be optimum. Further work on this subject may lead to useful result.

The optimum FWL-LQG regulator has been designed subject to the unity ℓ_2-scaling constraint. This approach however, ignores additional scaling that is generally required in order to handle initial condition type disturbance transients for which ℓ_∞-scaling is more appropriate. There are many considerations that contribute to the selection of scaling factors [Hwang (1975), Moroney et. al (1983)]. In practice, simulation studies would most likely aid this process. When greater dynamic range is required, for example to handle the transient dynamics or the tracking problem, the block-floating-point arithmetic [Oppenheim (1970), Heath et. al (1979) and Williamson et. al (1985)] can be used to replace the fixed point arithmetic. We have not explore the use of the block-floating-point arithmetic for control.

The coefficient wordlength consideration in the LQG regulator has also been addressed in chapter 5. We have derived an analytical expression of the degradation on the compensator performance due to a finite coefficient wordlength implementation. This result is similar to the result described in [Sasahara et. al (1984)] even though it was derived independently. Ours and theirs were derived based on the same source [Kawamata and Higuchi (1985)].

The round-off noise (due to the state quantization) analysis presented in

chapter 5 relies heavily on the validity of the additive noise model for round-off quantization. The model however, is no longer valid if the limit cycles produced by the quantization non-linearities exist and are significant. There are two types of limit cycles; namely the quantizer limit cycles and the overflow limit cycles [Kaiser (1976)]. In digital filtering, it has been shown that the influence of the overflow limit cycles on the performance is stronger than that of the quantizer limit cycles. However, the limit cycles may not be an important subject in the FWL-LQG design since the 'whiteness' of the process disturbance and of the measurement noise may obviate the occurrence of the limit cycles [Moroney (1983)]. In this thesis, we have not examined the effects of the limit cycles on the performance of the FWL-LQG regulators.

In chapter 6, we examined the implementation of a low complexity - low sensitivity compensator. The scaled single-input single-output (SISO) DRDF structure which was originally developed in digital filtering [Williamson (1987)] is a structure which has both the low complexity and the low sensitivity properties. We extended the procedures so that the scaled DRDF structure can be derived directly from the default structure. This is more convenient for control applications. The round-off noise and the coefficient sensitivity performance of the scaled SISO-DRDF structure has been examined and compared to the performance of other structures including the default, the direct form and the optimum (ie. the minimum round-off noise) structures.

The implementation of a multi-input multi-output (MIMO) DRDF structure was also examined in chapter 6. We showed that a simple extension of the procedures for deriving the scaled SISO-DRDF structure gives the MIMO version of the procedures. The FWL performance of the scaled MIMO-DRDF structure has also been investigated. As for the SISO-DRDF structure, the MIMO-DRDF structure requires only n_x (ie. the order of the compensator realization) more multiplication than would be required by the corresponding MIMO direct form structure. However, unlike the SISO-DRDF structure, the MIMO-DRDF representation is not unique,

and is determined by the selection of the controllability indices (described in lemma 6.4). In fact the freedom in selecting the controllability indices σ_i (and hence the MIMO-DRDF structure) would be beneficial for control applications to meet certain objectives. We have not explored this issue in this thesis.

APPENDIX A

LONGITUDINAL CONTROL OF A MODERN TRANSPORT AEROPLANE

A longitudinal control of a modern transport aeroplane which can be modelled as a 5–input 2–output 9^{th}–order continuous–time system is discussed in [Gangsaas et. al (1986)]. In this appendix, we consider a 6^{th}–order single–input single–output minimal system which is a simplified version of the above model. The model can be described by the following state space representation

$$\dot{x}(t) = Ax(t) + Bu(t) + \omega(t) \qquad (A.1)$$

$$y(t) = Cx(t) + \eta(t) \qquad (A.2)$$

where the state $x(t) \epsilon R^6$ are related to the physical variable as follows

$x_1(t)$ = 20.14*incremental forward velocity (ft/sec)

$x_2(t)$ = 1.19*angle of attack (deg)

$x_3(t)$ = 0.97*pitch rate (deg/s)

$x_4(t)$ = 2.66*pitch angle (deg)

$x_5(t)$ = 1.23*elevator position (deg)

$x_6(t)$ = 0.76*servo position (deg)

where * denotes times and where $x_i(t)$ is the i^{th} element of the state vector x(t). The control input $u(t) \epsilon R^1$ in (A.1) is the elevator command (deg) and the control output $y(t) \epsilon R^1$ is the pitch rate (deg/s). The matrices A, B and C are given below. The system matrix A is a (6x6) matrix

$$\begin{array}{cccccc}
-7.0200D-03 & 6.3390D-01 & 5.1800D-03 & -5.5566D-01 & -6.1120D-02 & 0.0000D+00 \\
-1.6540D-02 & -3.8892D-01 & 1.0057D+00 & 5.9100D-03 & -4.6320D-02 & 0.0000D+00 \\
6.1000D-04 & 3.5210D-01 & -4.7381D-01 & 0.0000D+00 & 1.7826D+00 & 0.0000D+00 \\
0.0000D+00 & 0.0000D+00 & 1.0000D+00 & 0.0000D+00 & 0.0000D+00 & 0.0000D+00 \\
0.0000D+00 & 0.0000D+00 & 0.0000D+00 & 0.0000D+00 & -2.0000D+01 & 2.0000D+01 \\
0.0000D+00 & 0.0000D+00 & 0.0000D+00 & 0.0000D+00 & 0.0000D+00 & -3.0000D+01
\end{array} \qquad (A.3)$$

The input matrix B is a (6x1) column matrix

```
0.0000D+00
0.0000D+00
0.0000D+00
0.0000D+00
0.0000D+00
3.0000D+01
```
(A.4)

The output matrix C is a (1x6) row matrix

```
0.0000D+00  0.0000D+00  1.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
```
(A.5)

The continuous-time processes $\{\omega(t),\eta(t);-\infty<t<\infty\}$ in (A.1) and (A.2) are zero-mean 'white' noises having covariance

$$\xi\left\{\begin{bmatrix}\omega(t)\\ \eta(t)\end{bmatrix}[\omega'(t)\ \eta(t)]\right\} = \begin{bmatrix}\Omega_c & 0\\ 0 & \Lambda\end{bmatrix} \tag{A.6}$$

where the matrices Ω_c and Λ are given below.

The process disturbance covariance matrix Ω_c is a (6x6) matrix

```
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00
0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  0.0000D+00  1.0000D+00
```
(A.7)

The measurement noise covariance Λ is a scalar

```
1.0000d-06
```
(A.8)

Consider the following quadratic index J_c

$$J_c = \xi\{\lim_{m\to\infty} \frac{1}{2m}\int_{-m}^{m} [x'(t)Q_c x(t) + u'(t)R_c u(t)]dt \tag{A.9}$$

where $Q_c \geq 0$ and $R_c > 0$ are given below.

The state weighting matrix $Q_c=H'H$ where H is a (1X6) row matrix

2.5000D-05	5.8395D-04	-8.6000D-06	0.0000D+00	-7.0650D-05	0.0000D+00
5.8395D-04	1.3640D-02	-2.0088D-04	0.0000D+00	-1.6502D-03	0.0000D+00
-8.6000D-06	-2.0088D-04	2.9584D-06	0.0000D+00	2.4304D-05	0.0000D+00
0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00
-7.0650D-05	-1.6502D-03	2.4304D-05	0.0000D+00	1.9966D-04	0.0000D+00
0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00	0.0000D+00

(A.10)

Note that z(t)=Hx(t) represents the vertical acceleration of the aeroplane (g).

The control weighting factor R_c is a scalar

1.0000D-01 (A.11)

The objective of the control design is to control the pitch rate (ie. y(t) in (A.2) by using the elevator command u(t) in (A.1) such that the vertical acceleration z(t) is within the specified range. This goal can be achieved by minimizing the cost J_c in (A.9) subject to the system (A.1) and (A.2).

Assume the continuous-time model (A.1) and (A.2) is controlled digitally using a pulse-amplitude-modulation signal

$$u(t) = \sum_{k=-\infty}^{\infty} u(kT_c)p(t-kT_c) \tag{A.13}$$

where p(t) is given by

$$p(t) = \begin{cases} 1 \text{ for } t\epsilon(0,T_c) \\ 0 \text{ otherwise} \end{cases}$$

At the controlling instants $t_k=kT_c$, the discrete equivalent of the continuous-time system (A.1) and (A.2) can be written as follows

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \tag{A.14}$$

$$y(kT_c) = Lx(kT_c) + \eta(kT_c) \tag{A.15}$$

where the matrices Φ, Γ and L for T_c=0.02sec are given below.

The system matrix Φ is a (6x6) matrix

```
 COLUMNS    1 THRU    3
 9.998575191928187D-01   1.262832392420975D-02   1.191000808255865D-04
-3.293791514488163D-04   9.923199245726563D-01   1.994284604536209D-02
 1.098377263273564D-05   6.981745373531006D-03   9.906387553007530D-01
 1.138798833072769D-07   7.001759981106059D-05   1.990600519736295D-02
 0.000000000000000D+00   0.000000000000000D+00   0.000000000000000D+00
 0.000000000000000D+00   0.000000000000000D+00   0.000000000000000D+00

 COLUMNS    4 THRU    6
-1.111166458062174D-02  -1.010782255426319D-03  -1.769569198369643D-04
 1.195770234216512D-04  -4.469035193628426D-04  -9.609798611549906D-05
 3.505255189248466D-07   2.923427591276637D-02   5.131636829871492D-03
 1.000000002333288D+00   3.123459388245558D-04   3.717606488732084D-05
 0.000000000000000D+00   6.703200520287676D-01   2.430168171492070D-01
 0.000000000000000D+00   0.000000000000000D+00   5.488116434541641D-01
```

(A.16)

The input matrix Γ is a (6x1) column matrix

```
-3.839915395807987D-05
-2.310975457561880D-05
 1.115281946619625D-03
 5.856977982083945D-06
 8.666313082202546D-02
 4.511883565458359D-01
```

(A.17)

The output matrix L is a (1x6) row matrix

```
0.    0.    1.    0.    0.    0.
```

(A.18)

The covariance of discrete processes $\{\omega(kT_c),\eta(kT_c):-\infty<k<\infty\}$ is given by

$$\xi\left\{\begin{bmatrix}\omega(kT_c)\\ \eta(kT_c)\end{bmatrix}[\omega'(kT_c)\ \ \eta(kT_c)]\right\} = \begin{bmatrix}\Omega & 0\\ 0 & \Lambda/T_c\end{bmatrix} \tag{A.19}$$

The process disturbances covariance matrix Ω is a (6x6) matrix

```
 COLUMNS    1 THRU    3
 1.398231284685686D-09   8.161550072020208D-10  -4.059133266894668D-08
 8.161550072020208D-10   4.779181706824628D-10  -2.369454708806988D-08
-4.059133266894668D-08  -2.369454708806988D-08   1.178386983803121D-06
-2.380754748982317D-10  -1.374736192574475D-10   6.910299853192613D-09
-2.554633477077882D-06  -1.515628386707430D-06   7.418107917384904D-05
-8.282713923773451D-06  -5.029931706139769D-06   2.406015786291162D-04

 COLUMNS    4 THRU    6
-2.380754748982317D-10  -2.554633477077882D-06  -8.282713923773452D-06
-1.374736192574475D-10  -1.515628386707430D-06  -5.029931706139769D-06
 6.910299853192613D-09   7.418107917384904D-05   2.406015786291162D-04
 4.199796849526074D-11   4.110156855813953D-07   1.219166535856234D-06
 4.110156855813953D-07   5.148667149134108D-03   1.991296083478731D-02
 1.219166535856234D-06   1.991296083478731D-02   1.164676313514978D-01
```

(A.20)

The measurement noise covariance Λ/T_c

```
5.000000000000000D-05
```
(A.21)

The discrete-time version J_d of the continuous-time cost J_c in (A.9) for a sampling period T_c is given by

$$J_d = \xi\{\lim_{n\to\infty} \frac{1}{2n} \sum_{k=-n}^{n} [x'(kT_c)Q_d x(kT_c) + 2x'(kT_c)M_d u(kT_c) + u'(kT_c)R_d u(kT_c)]\} \quad (A.22)$$

where the weighting matrices $Q_d \geq 0$, M_d and $R_d > 0$ for T_c=0.02sec are given below. The state weighting matrix Q_d is a (6x6) matrix

```
  COLUMNS   1 THRU   3
 2.480425418590327D-05   5.795450158852120D-04  -2.715737402094943D-06
 5.795450158852120D-04   1.354092056451630D-02  -6.343883431272638D-05
-2.715737402094944D-06  -6.343883431272638D-05   7.542333277873513D-07
-1.035790975898226D-07  -2.420342772271535D-06   3.192548637048291D-09
-5.833652249594507D-05  -1.362991905532932D-03   7.254933601191388D-06
-1.019332532607271D-05  -2.381843161520287D-04   4.483083675938718D-07

  COLUMNS   4 THRU   6
-1.035790975898226D-07  -5.833652249594507D-05  -1.019332532607271D-05
-2.420342772271535D-06  -1.362991905532932D-03  -2.381843161520287D-04
 3.192548637048291D-09   7.254933601191388D-06   4.483083675938719D-07
 5.778380763830468D-10   2.281292255288405D-07   5.447171172379364D-08
 2.281292255288404D-07   1.388525080352684D-04   2.269640550618790D-05
 5.447171172379364D-08   2.269640550618790D-05   5.184785749088448D-06
```
(A.23)

The cross input-state weighting M_d is a (6x1) column matrix

```
-2.209336517913487D-06
-5.162804526723970D-05
-6.057243132533780D-09
 1.364874283500783D-08
 4.729837354755177D-06
 1.259437025139311D-06
```
(A.24)

The input weighting R_d is a scalar

```
1.000003372458192D-01
```
(A.25)

The compensator which minimizes the cost J_d in (A.22) subject to the system (A.14) and (A.15) is governed by

$$\hat{x}((k+1)T_c) = \Phi\hat{x}(kT_c) + \Gamma u(kT_c) + K(y(kT_c) - L\hat{x}(kT_c)) \quad (A.26)$$

$$u(kT_c) = -G\hat{x}(kT_c) \quad (A.27)$$

where the matrices Φ, Γ and L are respectively given by (A.16), (A.17) and (A.18) and the matrices K and G are respectively the Kalman filter and the controller gains. From the matrices Φ, L, Ω and Λ/T_c defined in (A.16), (A.18), (A.20) and (A.21) the Kalman filter gain K in (A.26) can be computed [Anderson and Moore (1979)].

The Kalman filter gain K is a (6x1) column matrix

```
-2.418443357297475D-01
 2.090575409804330D-02
 5.409415470129207D-01
 4.033125967721040D-02
 2.930452358122123D+00
 1.387992699992277D+00
```

(A.28)

From the matrices Φ, Γ, Q_d, M_d and R_d defined in (A.16), (A.17), (A.23), (A.24) and (A.25) respectively, the controller gain G in (A.27) can be calculated [Åström and Wittenmark (1984)].

The controller gain G is a (1x6) row matrix

(A.29)

```
 COLUMNS   1 THRU   3
1.071682758094060D-05   3.058222314139755D-01   4.567920267701158D-01

 COLUMNS   4 THRU   6
5.980342186072287D-02   3.976538276591681D-02   2.622786374454104D-02
```

APPENDIX B

AN EULER-BERNOULLI BEAM

In this appendix, we describe a mathematical model of a flexible simply supported (Euler-Bernoulli) beam [Bateman (1964)]. Consider the simply supported beam depicted in Fig.B.1

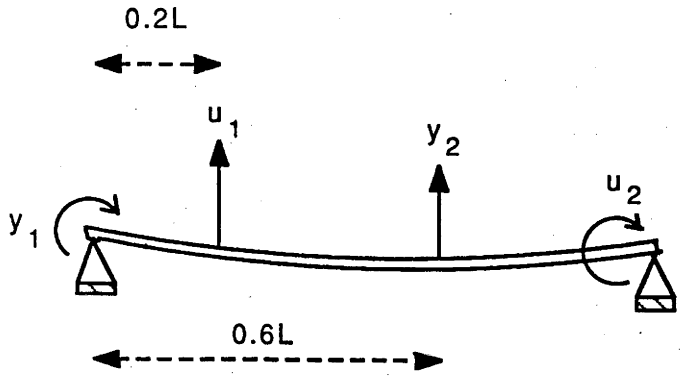


Fig.B.1 A simply supported beam

For simplicity, the shear deformation and rotary inertia effects are assumed to be negligible. This is the so called *Euler-Bernoulli beam*. The parameters of the beam are given by

a. The length of the beam $L=\pi$

b. The (uniform) mass density $\rho=2/L$

c. The modulus elasticity $EI=\rho$, where E and I are respectively Young's modulus and the moment of inertia of the cross section of the beam about a horizontal line.

The equations of motion of the beam can be described as follows

$$\ddot{d}_i(t) + 2\zeta w_i \dot{d}_i(t) + w_i^2 d_i(t) = b_i(u(t) + \omega(t)) \qquad (B.1)$$

where $\dot{d}_i(t)$ represents the first derivative of $d_i(t)$ with respect to t where $d_i(t)$ is the i^{th} mode deflection of the beam, w_i is the i^{th} natural frequency of the beam and ζ is damping ratio and where

$$\zeta = 0.005$$

$$w_i = (EI/\rho)^{\frac{1}{2}}(i\pi/L)^2 = i^2$$

$$b_i = [\psi_i(0.2L)\ \varphi(L)] \tag{B.2}$$

$$\Psi_i(0.2L) = \sin(0.2Li) \tag{B.3}$$

$$\varphi_i(L) = i\cos(Li) \tag{B.4}$$

and u(t) and $\omega(t)$ are respectively the actuator signal and the actuator noise. The actuator signal u(t) is a vector of the form

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$$

where $u_1(t)$ and $u_2(t)$ represent the force at location r_1=0.2L and the torque at location r_2=L, respectively. The output of the model are given by

$$y(t) = \sum_{i=1}^{N} c_i \dot{d}_i(t) \tag{B.5}$$

where N is the number of mode, $\dot{d}_i(t)$ is defined in (B.1) and c_i is given by

$$c_i = \begin{bmatrix} \varphi_i(0) \\ \Psi_i(0.6L) \end{bmatrix} \tag{B.6}$$

$$\varphi_i(0) = i \tag{B.7}$$

$$\psi_i(0.6L) = \sin(0.6Li) \tag{B.8}$$

The first 5 mode shapes of the beam are shown in Fig.B.2.

The state space representation of the Euler-Bernoulli beam for N=5 can be derived from the equations of motion (B.1) and (B.5), results in a 10^{th}-order 2-input 2-output continuous-time model described by

$$\dot{x}(t) = Ax(t) + B(u(t)+\omega(t)) \tag{B.9a}$$

$$y(t) = Cx(t) + \eta(t) \tag{B.9b}$$

where the state $x(t)\epsilon R^{10}$, the input $u(t)\epsilon R^2$ and the output $y(t)\epsilon R^2$.

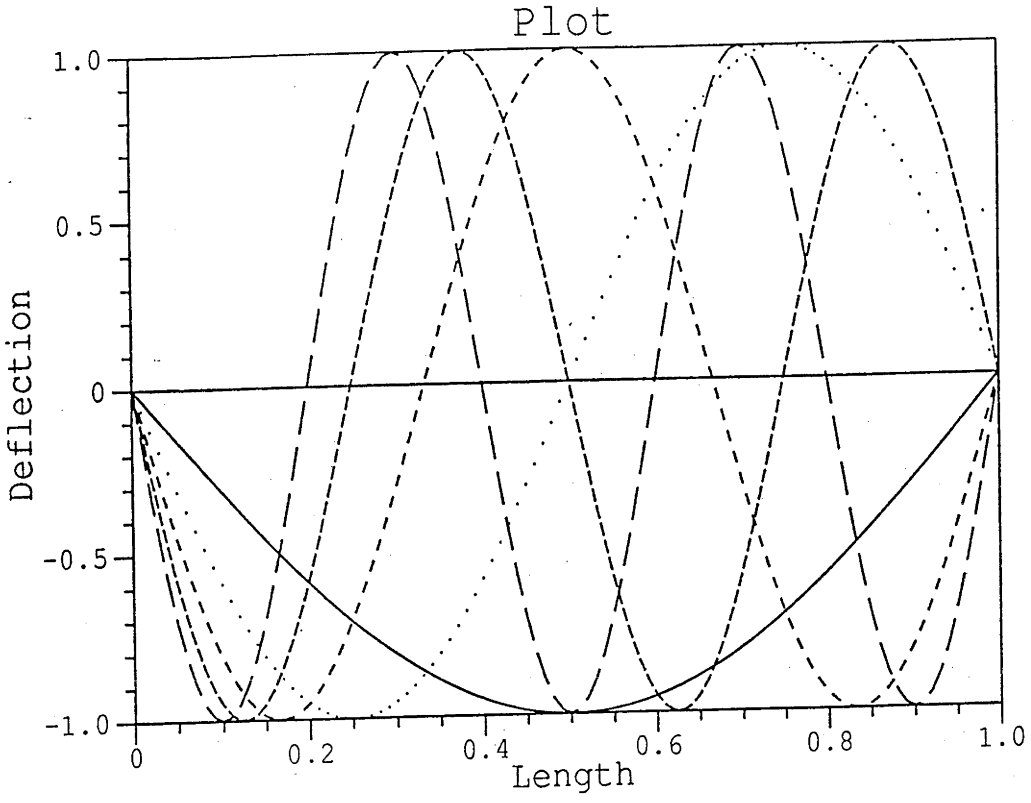


Fig.B.2 Mode shapes for the Euler-Bernoulli beam

The processes $\{\omega(t),\eta(t):-\infty<t<\infty\}$ in (B.9a-b) are assumed to be independent zero mean 'white' noise processes having covariance

$$\xi\left\{\begin{bmatrix}\omega(t)\\ \eta(t)\end{bmatrix}\begin{bmatrix}\omega'(t) & \eta'(t)\end{bmatrix}\right\} = \begin{bmatrix}\Omega_c & 0\\ 0 & \Lambda\end{bmatrix} \qquad (B.10)$$

where

$$\Omega_c = 10^{-3} I_2$$

$$\Lambda = 10^{-6} I_2$$

where I_2 is a (2x2) identity matrix. The matrices A, B and C in (B.9a-b) are given below.

The system matrix A is a (10x10) matrix of the form

$$A = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A_{55} \end{bmatrix} \tag{B.11}$$

where for i=1,2,...,5

$$A_{ii} = \begin{bmatrix} 0 & 1 \\ -w_i^2 & -2\zeta w_i \end{bmatrix} \tag{B.12}$$

where $w_i=i^2$ and $\zeta=0.005$.

The input matrix B is a (10x2) matrix of the form

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_5 \end{bmatrix} \tag{B.13}$$

where for i=1,2,...,5

$$B_i = \begin{bmatrix} 0 & 0 \\ \Psi_i(0.2L) & \varphi_i(L) \end{bmatrix} \tag{B.14}$$

where $\Psi_i(0.2L)$ and $\varphi_i(L)$ are defined in (B.3) and (B.4).

The output matrix C is a (2x10) matrix of the form

$$C = [C_1 \quad C_2 \quad \cdots \quad C_5] \tag{B.15}$$

where for i=1,2,...,5

$$C_i = \begin{bmatrix} \varphi_i(0) & 0 \\ \Psi_i(0.6L) & 0 \end{bmatrix} \tag{B.16}$$

where $\varphi_i(0)$ and $\Psi_i(0.6L)$ are defined in (B.7) and (B.8).

The purpose of the control design is to minimize the variance of output y(t)

$$\xi\{y(t)y'(t)\} \tag{B.17}$$

The criterion (B.17) can be restated as a continuous-time quadratic cost J_c as follows

$$J_c = \xi\{ \lim_{n\to\infty} \frac{1}{2n} \int_{-n}^{n} [x'(t)Q_c x(t) + u'(t)R_c u(t)]dt\} \tag{B.18}$$

with $Q_c=C'C$ where C is defined in (B.15) and $R_c=0.005I_2$ where I_2 is a (2x2) identity matrix.

The discrete equivalent of the continuous-time system (B.9) for a sampling period T_c=0.02sec is given by

$$x((k+1)T_c) = \Phi x(kT_c) + \Gamma u(kT_c) + \omega(kT_c) \tag{B.19}$$

$$y(kT_c) = Lx(kT_c) + \eta(kT_c) \tag{B.20}$$

where the matrices Φ, Γ and L are given below.

The system matrix Φ is a (10x10) matrix of the form

$$\Phi = \begin{bmatrix} \Phi_{11} & 0 & \cdots & 0 \\ 0 & \Phi_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \Phi_{55} \end{bmatrix} \tag{B.21}$$

Φ_{11} =

```
 9.998000200076504D-01   1.999666651312637D-02
-1.999666651312637D-02   9.996000533425191D-01
```

Φ_{22} =

```
 9.992002133356609D-01   1.999066824655187D-02
-7.996267298620749D-02   9.988003999707299D-01
```

Φ_{33} =

```
 9.982008998320406D-01   1.998200651237654D-02
-1.798380586113889D-01   9.976014396366693D-01
```

Φ_{44} =

```
 9.968025590622555D-01   1.997068370850761D-02
-3.195309393361218D-01   9.960037317139152D-01
```

Φ_{55} =

```
 9.950058300854875D-01   1.995670287084840D-02
-4.989175717712101D-01   9.940079949419451D-01
```

The input matrix Γ is a (10x2) matrix

```
 1.975179067960252D-04   1.617872123458151D-04
 1.975047436578414D-02   1.617764303996765D-02
 6.178691778880770D-05   1.901605796914142D-04
 6.177456217096160D-03   1.901225530097777D-02
-1.781122208133551D-04   6.177250294247453D-05
-1.780409816890196D-02   6.174779594035226D-03
-1.174630392676526D-04  -1.174630392676526D-04
-1.173847336205833D-02  -1.173847336205833D-02
 1.412564565179845D-04  -1.997667965804991D-04
 1.411151993010195D-02  -1.995670287084840D-02
```

(B.22)

The output matrix L is a (2x10) matrix

```
 COLUMNS   1 THRU   3
0.000000000000000D+00   9.876883405951377D-01   0.000000000000000D+00
0.000000000000000D+00   8.090169943749474D-01   0.000000000000000D+00

 COLUMNS   4 THRU   6
3.090169943749475D-01   0.000000000000000D+00  -8.910065241883678D-01
9.510565162951536D-01   0.000000000000000D+00   3.090169943749474D-01

 COLUMNS   7 THRU   9
0.000000000000000D+00  -5.877852522924732D-01   0.000000000000000D+00
0.000000000000000D+00  -5.877852522924732D-01   0.000000000000000D+00

 COLUMNS  10 THRU  10
 7.071067811865474D-01
-1.000000000000000D+00
```

(B.23)

The covariance of the discrete processes $\{\omega(kT_c),\eta(kT_c):-\infty<k<\infty\}$ is given by

$$\xi\left\{\begin{bmatrix}\omega(kT_c)\\ \eta(kT_c)\end{bmatrix}\begin{bmatrix}\omega'(kT_c) & \eta'(kT_c)\end{bmatrix}\right\} = \begin{bmatrix}\Omega & 0\\ 0 & \Lambda/T_c\end{bmatrix} \tag{B.24}$$

where the covariance Ω and Λ/T_c are given below.

The covariance Ω is a (10x10) matrix

```
 COLUMNS    1 THRU    3
 4.345765056674710D-09   3.258987005567606D-07   2.864471344363681D-09
 3.258987005567606D-07   3.258987049845556D-05   2.148192400298760D-07
 2.864471344363681D-09   2.148192400298760D-07   2.665013896771185D-09
 2.147619494573064D-07   2.147906035342722D-05   1.998134173917193D-07
-1.678922374346313D-09  -1.259149838279669D-07   4.943964795850926D-11
-1.258310317767546D-07  -1.258730181317791D-05   3.705483067341557D-09
-2.813235522256728D-09  -2.109972482051503D-07  -1.972606417982659D-09
-2.107440197259015D-07  -2.108706628570612D-05  -1.477753316679104D-07
-2.945364546911666D-10  -2.209222502079822D-08  -1.950173844499692D-09
-2.205097960479231D-08  -2.207160685505602D-06  -1.460072793374379D-07

 COLUMNS    4 THRU    6
 2.147619494573064D-07  -1.678922374346313D-09  -1.258310317767546D-07
 2.147906035342722D-05  -1.259149838279669D-07  -1.258730181317791D-05
 1.998134173917193D-07   4.943964795850926D-11   3.705483067341557D-09
 1.998134420934521D-05   3.706966409222764D-09   3.706225658002997D-07
 3.706966409222764D-09   2.368917869236592D-09   1.775569152416568D-07
 3.706225658002997D-07   1.775569152416568D-07   1.775570037322217D-05
-1.479134391729134D-07   9.108362694294893D-10   6.827344728089649D-08
-1.478444467130817D-05   6.823700511522067D-08   6.825528304361811D-06
-1.462413671100774D-07  -2.499240127968470D-09  -1.873480768799502D-07
-1.461244142528912D-05  -1.871230676036284D-07  -1.872358066009965D-05

 COLUMNS    7 THRU    9
-2.813235522256728D-09  -2.107440197259015D-07  -2.945364546911666D-10
-2.109972482051503D-07  -2.108706628570612D-05  -2.209222502079822D-08
-1.972606417982659D-09  -1.477753316679104D-07  -1.950173844499692D-09
-1.479134391729134D-07  -1.478444467130817D-05  -1.462413671100774D-07
 9.108362694294893D-10   6.823700511522067D-08  -2.499240127968470D-09
 6.827344728089649D-08   6.825528304361811D-06  -1.873480768799502D-07
 1.839160640502483D-09   1.377917630118586D-07   4.580273229581121D-10
 1.377917630118586D-07   1.377919548149001D-05   3.431821924778443D-08
 4.580273229581121D-10   3.431821924778443D-08   3.989015670731858D-09
 3.429530805400735D-08   3.430683542880547D-06   2.987025053995879D-07

 COLUMNS   10 THRU   10
-2.205097960479231D-08
-2.207160685505602D-06
-1.460072793374379D-07
-1.461244142528912D-05
-1.871230676036284D-07
-1.872358066009965D-05
 3.429530805400735D-08
 3.430683542880547D-06
 2.987025053995879D-07
 2.987034448171242D-05
```

(B.25)

The covariance Λ/T_c is a (2x2) matrix

```
 5.000000000000000D-05   0.000000000000000D+00
 0.000000000000000D+00   5.000000000000000D-05
```

(B.26)

The discrete version J_d of the continuous-time cost J_c defined in (B.18) is given by

$$J_d = \xi\{\lim_{m\to\infty} \frac{1}{2m} \sum_{k=-m}^{m} [x'(kT_c)Q_d x(kT_c) + x'(kT_c)M_d u(kT_c) + u'(kT_c)R_d u(kT_c)]\} \qquad (B.27)$$

where the matrices $Q_d \geq 0$, M_d and $R_d > 0$ for a sampling period T_c=0.02sec are given below.

The state weighting matrix Q_d is a (10x10) matrix

```
  COLUMNS   1 THRU   3
 2.172882528730039D-04  -1.629493502832209D-02   5.728942689761965D-04
-1.629493502832209D-02   1.629493524922729D+00  -4.296384800724893D-02
 5.728942689761965D-04  -4.296384800724893D-02   2.132011117801699D-03
-1.073809747318249D-02   1.073953017671287D+00  -3.996268347952186D-02
-7.555150685921328D-04   5.666174272425967D-02   8.899136634135893D-05
 6.291551589021879D-03  -6.293650906588124D-01  -7.410966134899586D-04
-2.250588418210783D-03   1.687977985690876D-01  -6.312340538680750D-03
 1.053720098659962D-02  -1.054353314285075D+00   2.955506633443453D-02
-3.681705684301546D-04   2.761528127680606D-02  -9.750869224250400D-03
 1.102548980270967D-03  -1.103580342752436D-01   2.920145586831621D-02

  COLUMNS   4 THRU   6
-1.073809747318249D-02  -7.555150685921328D-04   6.291551589021879D-03
 1.073953017671287D+00   5.666174272425967D-02  -6.293650906588124D-01
-3.996268347952186D-02   8.899136634135893D-05  -7.410966134899586D-04
 9.990672104671522D-01  -1.668134884199262D-03   1.853112829001181D-02
-1.668134884199262D-03   9.594117372135822D-03  -7.990061186107204D-02
 1.853112829001181D-02  -7.990061186107204D-02   8.877850186609008D-01
 1.833075134179310-01   6.558021141071501D-03  -5.461875782630068D-02
-7.392222335652179D-01  -3.070665230273212D-02   3.412764152179812D-01
 1.828017088929165D-01  -2.811645144469145D-02   2.341850961066901D-01
-7.306220712641866D-01   8.420538042401447D-02  -9.361790330045788D-01

  COLUMNS   7 THRU   9
-2.250588418210783D-03   1.053720098659962D-02  -3.681705684301546D-04
 1.687977985690876D-01  -1.054353314285075D+00   2.761528127680606D-02
-6.312340538680750D-03   2.955506633443453D-02  -9.750869224250400D-03
 1.833075134179310-01  -7.392222335652179D-01   1.828017088929165D-01
 6.558021141071501D-03  -3.070665230273212D-02  -2.811645144469145D-02
-5.461875782630068D-02   3.412764152179812D-01   2.341850961066901D-01
 2.354125620265843D-02  -1.102334104126423D-01   9.160546460803901D-03
-1.102334104126423D-01   6.889597740742203D-01  -4.289777406095171D-02
 9.160546460803901D-03  -4.289777406095171D-02   1.246567397326688D-01
-2.743624644397846D-02   1.715341771439387D-01  -3.733781317599401D-01

  COLUMNS  10 THRU  10
 1.102548980270967D-03
-1.103580342752436D-01
 2.920145586831621D-02
-7.306220712641866D-01
 8.420538042401447D-02
-9.361790330045788D-01
-2.743624644397846D-02
 1.715341771439387D-01
-3.733781317599401D-01
 1.493517224084686D+00
```

(B.28)

The cross state-input weighting matrix M_d is a (10x2) matrix

```
-4.059336110077539D-04  -3.834685591791344D-04
 3.044298616951835D-02   2.875831256044725D-02
-6.778352592003702D-04  -1.595381065690375D-03
 1.270468798503847D-02   2.990540702085110D-02
 2.725335634077435D-03  -6.230893715714900D-04
-2.269731883390004D-02   5.190536820829155D-03
 3.965509732419033D-03   4.327688378609372D-03
-1.856697186825218D-02  -2.026309883854693D-02
-4.855913129457697D-03   8.904444487730125D-03
 1.454429694127755D-02  -2.666928906619792D-02
```

(B.29)

The input weighting matrix R_d is a (2x2) matrix

```
1.506136974430462D-03   3.474398305100332D-04
3.474398305100332D-04   1.726113124807569D-03
```

(B.30)

The comepnsator which minimizes the cost J_d in (B.27) subject to the system (B.19) and (B.20) is given by

$$\hat{x}((k+1)T_c) = \Phi\hat{x}(kT_c) + \Gamma u(kT_c) + K(y(kT_c) - L\hat{x}(kT_c)) \qquad (B.31)$$

$$u(kT_c) = - G\hat{x}(kT_c) \qquad (B.32)$$

where the matrices Φ, Γ and L are respectively given by (B.21), (B.22) and (B.23) and where the matrices K and G are the Kalman filter and controller gains. From the matrices Φ, L, Ω and Λ/T_c defined in (B.21), (B.22), (B.25) and (B.26), the Kalman filter gain K in (B.31) can be computed [Anderson and Moore (1979)].

The Kalman filter gain K is a (10x2) matrix

```
 4.576866492528533D-03   4.186236378973598D-03
 2.653386553357453D-01   1.922164403560620D-01
 3.044464592390919D-03   4.615462836126575D-03
 6.168155433356739D-02   2.513393602694527D-01
-5.798099619408852D-03   1.568803172375010D-03
-2.702678594959814D-01   1.115228025814145D-01
-4.350097425628652D-03  -3.887545713570310D-03
-1.538147051440313D-01  -1.421467436408278D-01
 6.049755644241722D-03  -7.617452921267522D-03
 2.356339932474449D-01  -2.930790602644586D-01
```

(B.33)

From the matrices Φ, Γ in (B.21) and (B.22) and from the weighting matrices Q_d, M_d and R_d in (B.28), (B.29) and (B.30), the controller gain G in (B.32) can be computed.

The controller gain G is a (2x10) matrix

```
  COLUMNS   1 THRU   3
-1.243784652850972D-01   1.676806277522931D+01  -5.435494155014574D-01
-1.493649821728386D-01   1.211138931132323D+01  -5.163363884228726D-01

  COLUMNS   4 THRU   6
 3.881124321429097D+00   1.731154408160750D+00  -1.714171798618005D+01
 1.588650122309704D+01  -3.401927372214995D-01   7.090897027609800D+00

  COLUMNS   7 THRU   9
 2.529298977798672D+00  -9.754020132203742D+00  -5.489031565167069D+00
 2.265380062024229D+00  -9.015262862777378D+00   6.837490902447975D+00

  COLUMNS  10 THRU  10
 1.503512044521812D+01
-1.866246262721986D+01
```

(B.34)

PUBLICATIONS ARISING FROM THE RESEARCH

Kadiman K. and D. Williamson (1987), "Discrete minimax linear quadratic regulation of continuous-time systems", *Automatica*, November (to appear).

Kadiman K. and D. Williamson (1987), "Multirate digital regulation of continuous-time systems", (submitted for publication).

Williamson D. and K. Kadiman (1985), "Minimax variance regulation of continuous-time systems", *Proc. American Control Conf.*, ACC-85, Boston, USA.

Williamson D. and K. Kadiman (1985), "Cyclostationarity in the digital regulation of continuous-time systems", 7th Int. Symp. of The Math. Theory of Networks and Systems, Stockholm, Sweden. In Byrnes C.I. and A. Linquist (Eds.), *Frequency Domain and State-space Methods for Linear Systems,* North-Holland, Amsterdam, pp297-309.

Williamson D. and K. Kadiman (1987), "Finite wordlength linear quadratic regulation", *Proc. Int. Symp. on Ccts. Systs.*, Philadelphia, USA.

Williamson D. and K. Kadiman (1988), "Numerically robust state-space digital controllers", *Proc. American Control Conf.*, ACC-88, Georgia, USA (to appear).

Williamson D. and K. Kadiman (1988), "Design and implementation of finite wordlength linear quadratic regulators", (under preparation).

REFERENCES

Abu-El-Haija A.I. and A.M. Peterson, (1979) "An approach to eliminate round-off errors in digital filters", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-27, pp195-198.

Ackermann J., (1985) *Sampled Data Control Systems*, Springer Verlag, Berlin.

Agarwal R.C. and C.S. Burrus (1975), "New recursive digital filter structures having low sensitivity and round-off noise", *IEEE Trans. Ccts. Systs.*, CAS-22, No.12, pp921-927.

Ahmed M.E. and P.R. Belanger, (1984a) "Scaling and round-off in fixed-point implementation of control algorithms", *IEEE Trans. Ind. Electron.*, IE-31, pp228-234.

Anderson B.D.O. and J.B. Moore, (1979) *Optimal Filtering*, Englewood Cliffs, New Jersey, Prentice-Hall, Inc.

Antoniou A., C. Charalambous and Z. Motamedi, (1983) "Two methods for the reduction of quantization effects in recursive digital filters", *IEEE Trans. Ccts. Syst.*, CAS-30, pp160-167.

Åström K.J., (1970) *Introduction to Stochastic Control Theory*, Academic Press.

Åström K.J. and B. Wittenmark, (1984) *Computer Controlled Systems*, Prentice-Hall, Englewood Cliffs, New Jersey.

Åström K.J., P. Hagander and J. Sternby, (1984) "Zeros of sampled systems", *Automatica*, 20, pp31-38.

Åström K.J., V. Borisson, L. Ljung and B. Wittenmark (1977), "Theory and applications of self-tuning regulators", *Automatica*, Vol.13, pp457-476.

Athans M., guest ed., (1971) *IEEE Transactions Auto. Control, Special Issue on Linear-Quadratic-Gaussian Problem*, AC-16.

Avenhaus E., (1972) "On the design of digital filters with coefficients of limited wordlength", *IEEE Trans. Audio Electroacaustics*, AU-20, pp206-212.

Barnes C.W., B.N. Tran and S.H. Leung, (1985) "On the statistics of fixed-point round-off error", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-33, pp595-606.

Barnes C.W., (1979) "Round-off noise and overflow in normal digital filters", *IEEE Trans. Ccts. Systs.*, CAS-26, pp154-159.

Barnes C.W., (1984a) "On the design of optimal state-space realizations of second-order digital filters", *IEEE Trans. Ccts. Systs.*, CAS-31, pp602-608.

Barnes C.W., (1984b) "Computationally efficient second-order digital filter sections with low round-off noise gain", *IEEE Trans. Ccts. Systs.*, CAS-31, pp841-847.

Barnes C.W. and A.T. Fam (1977), "Minimum norm recursive digital filters that are free of overflow limit cycles", *IEEE Trans. Ccts. Systs.*, CAS-24, pp569-574.

Bateman H., (1964) *Partial Differential Equations of Mathematical Physics,* Cambridge University Press.

Bellman R., (1970) *Introduction to matrix analysis*, McGraw Hill, New York.

Bendat J.S and A.G. Piersol, (1966). *Measurement and Analysis of Random Data*, John Wiley, New York.

Bennet W.R., (1958) "Statistics of regenerative digital transmission", *Bell Syst. Tech. Journal*, Vol.37, pp1501-1542.

Bertram J.E., (1958) "The effects of quantization in sampled feedback systems", *Trans. Amer. Inst. Elec. Engrs.*, Vol.77, Pt.2, pp177-182.

Bitmead R.R., A.C. Tsoi and P.J. Parker, (1986) "A Kalman filtering approach to short-time Fourier analysis", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-34, pp1493-1501.

Bonzanigo F., (1974) "Comment on 'Round-off noise and attenuation sensitivity in digital filters with fixed-point arithmetic", *IEEE Trans. Ccts. Syst.*, CAS-21, pp809-810.

Brockett R.W., (1970) *Finite Dimensional Linear Systems*, J. Wiley and Sons, Inc., New York.

Broussard J.R. and D.P. Glasson, (1980) "Optimal multirate flight control design", *Proc. Joint Auto. Control Conf.*, San Francisco, WPI-E.

Broussard J.R., D.R. Downing and W.H. Bryant, (1985) "Design and flight testing of a digital optimal control general aviation autopilot", *Automatica*, 21, pp23-34.

Chan D.S.K. and L.R. Rabiner, (1973) "Analysis of quantization errors in the direct form for finite impulse response digital filters", *IEEE Trans. Audio & Electroacoustics*, AU-21, pp354-366.

Chan S.W., G.C. Goodwin and K.S. Sin, (1984) "Convergence properties of the Riccati difference equation in optimal filtering of non-stabilizable systems", *IEEE Trans. Auto. Contr.*, AC-29, pp110-118.

Chen C.T., (1984) *Linear System Theory and Design*, Holt, Rinehart and Winston, Holt-Saunders, Japan.

Claasen T.A.C.M., W.F.G. Mecklenbräuker and J.B.H. Peek, (1976) "Effects of quantization and overflow in recursive digital filters", *IEEE Trans Acoust. Speech Sig. Process.*, Vol.ASSP-24, pp517-529.

Crochiere R.E., (1975) "A new statistical approach to the coefficient wordlength problem for digital filters", *IEEE Trans. Ccts. Systs.*, CAS-22, pp190-196.

Curry E.E., (1967) "The analysis of round-off and truncation errors in a hybrid control system", *IEEE Trans. Auto. Control*, AC-13, pp601-604.

de Souza C.E. and G.C. Goodwin, (1984) "Intersample variances in discrete minimum variance control", *IEEE Trans. Auto. Control*, AC-29, pp759-761.

Deutsch R., (1954) "Detection of modulated noise-like signals", *IEEE Trans. Info. Theory*, pp.106-122.

Doyle J.C. and G. Stein, (1981) "Multivariable feedback design: concepts for classical/modern synthesis", *IEEE Trans. Auto. Control*, AC-26, pp4-16.

Fettweis A., (1972) "On the connection between multiplier wordlength limitations and round-off noise in digital filters", *IEEE Trans. Circiut Theory*, CT-19, pp486-491.

Fettweis A., (1973) "Round-off noise and attenuation sensitivity in digital filters with fixed-point arithmetic", *IEEE Trans. Circuit Theory*, CT-20, pp174-175.

Fettweis A., (1974) "On properties of floating-point round-off noise", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-22, pp149-151.

Fletcher R., (1980) *Practical Methods of Optimization*, John Wiley & Son, Vol.1&2.

Franklin G.F. and J.D. Powell, (1980) *Digital Control of Dynamic Systems*, Reading, Mass., Addison-Wesley.

Franks L., (1969) *Signal Theory*, Prentice Hall.

Gangsaas D., K.R. Bruce, J.D. Blight and U.L. Ly, (1986) "Application of modern synthesis to aircraft control : Three case studies", *IEEE Trans. Auto. Control*, AC-31, pp995-1014.

Gardner W.A., (1972) *Representation and Estimation of cyclostationary processes*, Ph.D dissertation, Uni. Massachusetts, Amherst.

Gardner W.A. and L.E. Franks, (1975) "Characterization of cyclostationary random signal processes", *IEEE Trans. on Info. Theory*, IT-21, pp4-14.

Gelb A., ed., (1974) *Applied Optimal Estimation*, MIT Press, Cambridge, Mass.

Gerheim A.P., (1984) "Numerical solution of Lyapunov equation for narrow band digital filters", *IEEE Trans. Ccts Systs.* CAS-31, pp991-992.

Gill P.E., W. Murray and M.H. Wright, (1981) *Practical Optimization*, Academic Press.

Glasson D.P., (1983) "Development and application of multirate digital control", *Cont. Syst. Magazine*, November, pp1-8.

Glover K., (1984) "All optimal Hankel-norm approximation of linear multivariable systems and their L^{∞}-error bounds", *Int. J. Cont.*, 39, pp1115-1193.

Goodwin G.C., R.J. Evans, R.L. Leal and R.A. Feik, (1986) "Sinusoidal Disturbance Rejection with Application to Helicopter Flight Data Estimation", *IEEE Trans. Acoust. Speech and Sig. Process.*, ASSP-34, pp479-484.

Goodwin G.C. and K.S. Sin, (1984) *Adaptive Filtering, Prediction and Control,* Prentice Hall, Englewood Cliffs, NJ.

Goff K.W., (1966) "A Systematic approach to DDC design", *ISA Journal,* December.

Grimble M.J. and R.J. Patton, (1980) "The design of dynamic ship positioning control systems using stochastic opimal control theory", *Optimal Control Application and Methods*, Vol.1, pp167-202.

Gupta A. and H.D. Toong, (1983), "Microprocessors : the first twelve years", *Proc. IEEE*, 71, pp1236-1256.

Gupta A. and H.D. Toong, (1984), "Microcomputers in industrial control applications", *IEEE Trans. Ind. Electron.*, IE-31, pp109-119.

Hall E.L., D.D. Lynch and S.J. Dwyer (1970), "Generation of products and quotients using approximate binary logarithms for digital filter applications", *IEEE Trans. Comput.*, C-19, pp97-102

Heath J.R., H.T. Nagle and S.G. Shiva, (1979) "Realization of digital filters using input scaled floating-point arithmetic", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-27, pp469-477.

Hwang S.Y., (1974) "On optimization of cascade fixed-point digital filters", *IEEE Trans. Ccts. Systs.*, CAS-21, pp163-165.

Hwang S.Y., (1975a) "Dynamic range constraints in state-space digital filtering", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-23, pp591-593.

Hwang S.Y., (1975b) "On monotonicity of Lp and ℓp norms" *IEEE Trans. Acouts. Speech Sig. Process.*, ASSP-23, pp593-594

Hwang S.Y., (1976) "Round-off noise in state-space digital filtering: a general analysis", *IEEE Trans. Acoust. Speech Sig. Processing*, ASSP-24, pp256-262.

Hwang S.Y., (1977) "Minimum uncorrelated unit noise in state-space digital filters", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-25, pp273-281.

Jackson L.B., (1970a) "On the interaction of round-off noise and dynamic range in digital filters", *Bell Syst. Tech. Journal*, Vol.49, No.2, pp159-184.

Jackson L.B., (1970b) "Round-off noise analysis for fixed-point digital filters realized in cascade or parallel form", *IEEE Trans. Audio Electroacoustics*, AU-18, pp107-122.

Jackson L.B., (1976) "Round-off noise bounds derived from coefficient sensitivites for digital filters", *IEEE Trans. Ccts. Systs.*, CAS-23, pp481-485.

Jackson L.B., A.G. Lindgren and Y. Kim, (1979) "Optimal synthesis of second order state-space structures for digital filters", *IEEE Trans. Ccts. Systs.*, CAS-26, pp149-153.

Jenkins W.K., (1979) "Recent advances in residue number techniques for reqursive digital filtering" *IEEE Trans. Acoust. Speech SIg. Process.*, ASSP-27, pp19-30.

Jing Z. and A.T. Fam, (1986) "A new scheme for designing IIR filters with finite wordlength coefficients", *IEEE Trans. Acoust., Speech Sig. Process.*, ASSP-34, pp1335-1336.

Johnson G.W., (1965) "Upper bound on dynamic quantization error in digital control systems via the direct method of Lyapunov", *IEEE Trans. Auto. Control*, AC-10, pp439-448.

Jordan K.L., (1961) *Discrete Representations of Random Signals*, PhD Dissertation, MIT, Cambridge.

Kadiman K. and D. Williamson (1987), "Discrete minimax linear quadratic regulation of continuous-time systems", *Automatica*, November.

Kadiman K. and D. Williamson (1987), "Multirate digital regulation of continuous-time systems", (submitted for publication).

Kailath T., (1974) "A view of three decades of linear filtering theory," *IEEE Trans. Inform. Theory,* IT-20, pp146-181.

Kailath T., (1980) *Linear Systems*, Englewood Cliffs, New Jersey, Prentice Hall.

Kaiser J.F, (1976) "On the limit cycles problem", *Proc. IEEE Inter. Conf. Acoust. Speech. Sig. Process.*, pp642-644.

Kalman R.E, (1960), "A new approach to linear filtering and prediction problem", *J. Basic Eng., Trans. ASME,* series D, vol.82, no.1, pp35.45.

Kalman R.E., (1963) "New methods in Wiener filtering theory", *Proc. Symp. Eng. Appl. Random Functions Theory and Probability* (eds. J.L. Bogdanoff and F.Kozin), John Wiley & Sons, NY.

Kalman R.E. and R.S. Bucy, (1961) "New Results in linear filtering and prediction theory", *J. Basic Eng. Trans. ASME*, series D, vol.83, no.3, pp95-108.

Kalman R.E., (1972) "Kronecker invariants and feedback", in *Ordinary Differential Equations,* L. Weiss, ed., Academic Press, New York, pp459-471.

Katz P., (1981) *Digital Control using Microprocessors,* Prentice Hall, Englewood Cliffs, New Jersey.

Kawamata M. and T. Higuchi, (1985) "A unified approach to the optimal synthesis of fixed-point state space digital filters", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-33, pp911-920.

Knowles J.B. and R. Edwards, (1965) "Effect of a finite wordlength computer in a sampled data feedback system", *Proc. IEEE.* Vol.112, No.6, pp1197-1207.

Knowles J.B. and R. Edwards, (1966) "Computational error effects in a direct digital control system", *Automatica*, 4, pp7-15.

Knowles J.B. and E.M. Olcayto, (1968) "Coefficient accuracy and digital filter response", *IEEE Trans. Ccts. Systs.*, CAS-15, pp31-41.

Kolmogorov A.N., (1941) "Interpolation and Extrapolation", *Bull. de l'académic des sciences de U.S.S.R*, Ser. Math.5, pp3–14.

Kucera V., (1972) "A Contribution to amtrix quadratic equations", *IEEE Trans. Auto. Control*, AC–17, pp344–347.

Kuo B.C., (1963) *Analisys and Synthesis of sampled–data control systems*, Prentice Hall, Englewood Cliffs, NJ.

Kuo B.C., (1980) *Digital Control Systems*, Holt, Reinhart and Winston, Tokyo.

Kuo B.C. and D.W. Peterson, (1973) "Optimal discretization of continuous–data control system. *Automatica,* vol.9, pp125–129.

Kwakernaak H. and R. Sivan, (1972) *Linear Optimal Control Systems*, J.Wiley & Sons, New York.

Lamaire R.O. and J.H. Lang, (1986) "Performance of digital linear Regulators which use logarithmic arithmetic", IEEE Trans. Auto. Control AC–31, pp394–400.

Lang J.H., (1984) "On the design of a special–purpose digital control processor". *IEEE Trans. Auto. Control*, AC–29, pp195–201.

Lo P.H. and Y.C. Jeng, (1982) "Minimum sensitivity realization of second order recursive digital filter", *IEEE Trans. Acoust. Speech Sig. Process.,* ASSP–30, pp930–937.

Luenberger D.G., (1967) "Canonical forms for linear multivariable systems", *IEEE Trans. Auto. Control*, AC–28, pp290–293.

Martenson K., (1971) "On the matrix Riccati equation", *Inf. Sci.*, vol–3, pp17–49.

Mäkilä P.M., T. Westerlund and H.T. Toivonen, (1984) "Constrained linear quadratic gaussian control with process applications", *Automatica*, Vol.20, pp15–29.

Middleton R.H. and G.C. Goodwin, (1986) "Improved finite wordlength characteristics in digital control using delta operators", *IEEE Trans. Auto. Control*, AC–31, pp1015–1021.

Mita T., (1985) "Optimal digital feedback control systems counting computation time of control laws", *IEEE Trans. Auto. Control*, AC-30, pp542-548.

Mitchell E.E. and R. Demoyer, (1985) "A versatile digital controller algorithm incorporating a state observer and state feedback", *IEEE Trans. Ind. Electron.*, IE-32, pp78-84.

Moore B.C., (1981) "Principal component analysis in linear systems : controllability, observability and model reduction", *IEEE Trans. Auto. Control*, AC-26, pp17-32.

Moroney P., (1983) *Issues in the Implementation of Digital Feedback Compensators*, MIT Press, Cambridge, Massachusetts.

Moroney P., A.S. Willsky and P.K. Houpt, (1980) "The digital implementation of control compensators : the coefficient word-length issue", *IEEE Trans. Auto. Control*, AC-25, pp621-630.

Moroney P., A.S. Willsky and P.K. Houpt, (1983) "Round-off noise and scaling in the digital implementation of control compensators", *IEEE Trans Acoust. Speech Sig. Process.*, ASSP-31, pp1464-1477.

Mullis C.T. and R.A. Roberts, (1976a) "Synthesis of minimum roudoff noise fixed-point digital filters", *IEEE Trans. Ccts. Systs.*, CAS-23, pp551-562.

Mullis C.T. and R.A. Roberts, (1976b) "Round-off noise in digital filters-frequency transformations and invariants", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-24, pp538-550.

Mullis C.T. and R.A. Roberts, (1982) "An interpretation of error spectrum shaping in digital filters", *IEEE Trans. Acoust. Speech Sig. Process.*, ASSP-30, pp1013-1015.

Munson Jr D.C. and B. Liu, (1981) "Narrow band recursive filters with error spectrum shaping", *IEEE Trans. Ccts. Systs.*, CAS-28, pp160-163.

Nishimura S., K. Hirano and R.N. Pal, (1981) "A new class of very low sensitivity and low round-off noise recursive digital filters structures", *IEEE Trans. Ccts. Systs.*, CAS-28, pp1152-1158.

Ogura H., (1971) "Spectral representation of a periodic nonstationary random process", *IEEE Transaction on Information Theory*, IT-17, pp143-149.

Oppenheim A.V., (1970) "Realization of digital filters using block-floating-point arithmetic", *IEEE Trans. Audio. Electroacoust.*, AU-18, pp130-136.

Oppenheim A.V. and R.W. Schafer, (1975) *Digital Signal Processing*, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Orlandi G and G. Martinelli, (1984) "Low-sensitivity recursive digital filters obtained via the delay replacement", *IEEE Trans. Ccts. Systs.*, CAS-31, pp654-657.

Papoulis A., (1965) *Probability, Random Variables and Stochasits Processes,* McGraw Hill.

Phillips C.I., (1980) "Using simulation to calculate floating-point quantization errors", *Simulation*, pp207-214.

Phillips C.L. and H.T. Nagle, (1984) *Digital Control Systems Analysis and Design,* Prentice Hall, Englewood Cliffs, New Jersey.

Polak E., (1971) *Computational Methods in Optimization*, Academic Press.

Quinn W.J. and S.E. Williamson, (1985), "A frequency response method for predicting the level of intersample activity in discrete systems", *Int. J. Cont.*, 41, pp429-444.

Rabiner L.R. and B. Gold, (1975) *Theory and Application of Digital Signal Processing*, Prentice Hall Inc., Englewood Cliffs, New Jersey.

Rink R.E and H.Y. Chong, (1979) "Performance of state regulator systems with floating-point computation", *IEEE Trans. Auto. Control,* AC-24, pp411-421.

Sage A.P., (1960) *Optimal Systems Control*, Prentice Hall, Englewood Cliffs, New Jersey.

Sandberg I.W., (1967) "Floating-point round-off accumulation in digital filter realizations", *Bell Syst. Tech. J.*, vol.46, pp1775-1791.

Sandberg I.W. and J.F. Kaiser, (1972) "A bound on limit cycles in fixed-point implementations of digital filters", *IEEE Trans. Audio Electroacoust.*, AU-20, pp110-112.

Sasahara H., M. Kawamata and T. Higuchi, (1984) "Design of microprocessor-based LQG control systems with minimum quantization error", *Proc. IECON '84,* Tokyo, pp533-538.

Scharf L.L. and S. Sigurdsson, (1984) "Fixed-point implementation of fast Kalman predictors", *IEEE Trans. Auto. Control,* AC-29, pp850-852.

Slaughter J.B., (1964) "Quantization errors in digital control systems", *IEEE Trans. Auto. Control,* AC-9, pp70-74.

Soderstran M.A., (1977) "A high speed low cost recursive digital filter using residue number arithmetic", *Proc. IEEE,* 65, pp1065-1067.

Sridharan S., (1985) *Finite Wordlength Considerations in the Implementation of Digital Filters*, Ph.D Thesis, Uni. New South Wales, Sydney, Australia.

Sripad A.B, (1981) "Performance degradation in digitally implemented Kalman filters", *IEEE Trans. Aerospace Electron. Syst.*, AES-17, pp626-634.

Sripad A.B. and D.L. Snyder, (1977) "A necessary and sufficient condition for quantization errors to be uniform and white", *IEEE Trans. Acoust., Speech Sig. Process,* ASSP-25, pp442-448.

Stratonovich R.L., (1963) *"Topics in theory of Random Noise,* Gordon and Breach, UK.

Szcupak J. and S.K. Mitra, (1978) "On digital filter structures with low coefficient sensitivities", *Proc. IEEE,* Vol.66, pp1082-1083.

Tan C. and B.C. McInnis, (1982) "Adaptive digital control implemented using residue number systems", *IEEE Trans. Auto. Control*, AC-27, pp449-454.

Tao T.F., D. Bar Yehoshua and R. Martinez, (1977) "Applications of microprocessors in control problems", *Proc. 1977 Joint Control Conf.*, pp8-13.

Toivonen H.T., (1983) "Suboptimal control of linear discrete stochastic systems with linear input constraints", *IEEE Trans. Auto. Control*, AC-28, pp246-248.

Tou J.T., (1959) *Digital and sampled-data control systems*, McGraw Hill.

Troch I., (1973) "Sampling with arbitrary choice of the sampling instants", *Automatica*, 9, pp117-124.

Van Wingerden A.J.M. and W.L. de Koning, (1984) "The influence of finite wordlenth on digital optimal control", *IEEE Trans. Auto. Control*, AC-29, pp382-391.

Weinstein C. and A.V. Oppenheim, (1969) "A comparison of round-off noise in floating-point and fixed-point digital filter realizations", *Proc. IEEE*, vol.57, pp1181-1183.

Wiener N., (1949) *Extrapolation, interpolation and smoothing of stationary time-series*, MIT Press, Cambridge, Mass.

Williamson D., (1985) "Finite wordlength design of digital Kalman filters for state estimation", *IEEE Trans. Auto. Control*, AC-30, pp930-939.

Williamson D., (1986) "Minimum round-off noise and pole-zero sensitivity using integer residue feedback", *IEEE Trans Acoust. Speech Sig. Process.*, ASSP-34, pp1210-1220.

Williamson D., (1987a) "Sensitivity improvement in high order direct form structures for low pass narrow band filtering", *IEEE Trans. Acoust. Speech Sig. Process.*, (to appear).

Williamson D., (1987b) "Optimal Fourier analysis using finite precision fixed-point arithmetic" (submitted for publication)

Williamson D. and K. Kadiman (1985a), "Minimax variance regulation of continuous-time systems", *Proc. American Control Conf.*, ACC-85, Boston, USA.

Williamson D. and K. Kadiman (1985b), "Cyclostationarity in the digital regulation of continuous-time systems", 7th Int. Symp. of The Math. Theory of Networks and

Systems, Stockholm, Sweden. In Byrnes C.I.and A. Linquist (Eds.), *Frequency Domain and State-space Methods for Linear Systems,* North-Holland, Amsterdam, pp297-309.

Williamson D. and K. Kadiman (1987), "Finite wordlength linear quadratic regulation", *Proc. Int. Symp. on Ccts. Systs.,* Philadelphia, USA.

Williamson D. and K. Kadiman (1988), "Numerically robust state-space digital controllers", *Proc. American Control Conf.*, ACC-88, Georgia, USA (to appear).

Williamson D. and K. Kadiman (1987), "Design and implementation of finite wordlength linear quadratic regulators", (under preparation).

Williamson D. and S. Sridharan, (1985a) "An approach to coefficient wordlength reduction in digital fitlers", *IEEE Trans. Ccts. Systs.,* CAS-32, pp893-904.

Williamson D. and S. Sridharan, (1985b) "Residue feedback in digital filters using fractional feedback coefficients", *IEEE Trans. Acoust. Speech Sig. Process.,* ASSp-33, pp477-483.

Williamson D., S. Sridharan and P.G. McCrea, (1985) "A new approach for block floating-point arithmetic in recursive filters", *IEEE Trans. Ccts. Systs.,* CAS-32, pp719-729.

Willsky A.S., (1979) *Digital Signal Processing and Control and Estimation Theory-Points of Tangency, Areas of Intersection, and Parallel Directions,* MIT Press, Cambridge, Massachusetts.

Wolovich W.A. and P.L. Falb, (1969) "On the Structure of Multivariable Systems", *SIAM J. Control,* Vol.7, pp437-451.